Mathématiques et sciences humaines
Mathematics and social sciences

192 | Hiver 2010
Varia

Filiation de manuscrits sanskrits et arbres phylogénétiques

Filiation of Sanskrit manuscripts and phylogenetic trees

Marc Le Pouliquen

Édition électronique
URL : http://journals.openedition.org/msh/11919
DOI : 10.4000/msh.11919
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 15 décembre 2010
Pagination : 57-91
ISBN : 0987 6936
ISSN : 0987-6936

Référence électronique
Marc Le Pouliquen, « Filiation de manuscrits sanskrits et arbres phylogénétiques », *Mathématiques et sciences humaines* [En ligne], 192 | Hiver 2010, mis en ligne le 11 mars 2011, consulté le 02 mai 2019.
URL : http://journals.openedition.org/msh/11919 ; DOI : 10.4000/msh.11919

Math. & Sci. hum. / Mathematics and Social Sciences (48e année, n° 192, *2010*(4), p. 57–91)

FILIATION DE MANUSCRITS SANSKRITS ET ARBRES PHYLOGÉNÉTIQUES

Marc LE POULIQUEN[1]

RÉSUMÉ – *La fabrication d'un stemma codicum est l'une des approches les plus rigoureuses de la critique textuelle. Elle exige la reconstruction de l'histoire du texte en classifiant le corpus pour décider si un groupe de manuscrits est engendré par un intermédiaire perdu. Pour classifier notre corpus, nous employons des méthodes de l'analyse textuelle informatisée et de la reconstruction phylogénétique afin d'établir un arbre de la filiation. Les techniques employées sont dédiées à un corpus de manuscrits sanskrits avec toutes les spécificités de cette langue.*

MOTS CLÉS – Arbres phylogénétiques, Distance d'arbre, Filiation de manuscrits, Sanskrit

SUMMARY – Filiation of sanskrit manuscripts and phylogenetic trees
The establishment of a stemma codicum is one of the most rigorous approaches of textual criticism. It requires the rebuilding of the history of the text by classifying the corpus to decide if a group of manuscripts is generated by a lost intermediary. To cluster our corpus, we use methods of the computerized textual analysis and phylogenetic reconstruction in order to establish the tree of filiation or pedigree. The method employed has been developed in editing sanskrit manuscripts with all specificities of this language.

KEYWORDS – Filiation of manuscripts, Phylogenetic trees, Sanskrit, Tree metrics

1. DÉDICACE

Jean-Pierre Barthélémy nous a quitté le 21 juin 2010. Il fut mon directeur de thèse. Cet article est dédié à cet homme de culture et de science, qui m'a beaucoup appris.

2. INTRODUCTION

Dans cet article, nous nous intéressons à la construction du « stemma codicum » de manuscrits sanskrits qui est l'une des problématiques de l'édition critique de textes anciens.

[1]Département Logique des Usages, Sciences Sociales et de l'Information (LUSSI), Traitement Algorithmique et Matériel de la Communication, de l'Information et de la Connaissance (TAMCIC), UMR CNRS 2872, ENST Bretagne, BP 832, 29285 Brest Cedex, marc.lepouliquen@enst-bretagne.fr et IUP Génie Mécanique et Productique, Université de Bretagne Occidentale, 6 avenue Le Gorgeu, CS93837, 29238 Brest Cedex 3

Le corpus de textes anciens, sur lequel nous avons travaillé, est une tradition textuelle de manuscrits sanskrits sur le plus ancien commentaire de la grammaire de *Pāṇini* qui nous soit parvenu, la *kāśikāvṛtti* ou Glose de Bénarès (cf. Figure 1). Elle date probablement du VII^{e} siècle de notre ère et est attribuée aux auteurs *Jayāditya* et *Vāmana*. La grammaire de *Pāṇini* est connue comme une des premières grammaires et est constituée d'un ensemble de règles qui peuvent difficilement se comprendre sans les explications d'un commentaire tel que la *kāśikāvṛtti*. La Glose de Bénarès qui est le plus répandu et le plus pédagogique des commentaires de *Pāṇini* n'a jamais fait l'objet d'une édition critique ou d'une traduction complète.

L'édition critique reconstitue au mieux, à partir des différents manuscrits conservés, l'œuvre telle que l'auteur l'a voulue. Ce domaine n'est actuellement que très peu informatisé et ce, malgré de nombreuses tâches répétitives. Des tentatives d'utilisation de l'informatique commencent à voir le jour (*Collate* en Angleterre, *Tustep* en Allemagne, etc.) mais les logiciels ne sont pas beaucoup utilisés par les éditeurs. De plus, si ces logiciels donnent des résultats intéressants pour le latin et le grec, ils ne sont pas du tout adaptés aux problématiques des langues orientales comme le sanskrit.

L'édition critique des textes anciens peut se résumer ainsi:

- Inventaire des manuscrits du corpus.
- Études codicologiques[2] et paléographiques[3] afin d'effectuer un premier classement des manuscrits.
- Collation des textes.
- Élaboration du stemma codicum afin d'expliquer l'histoire du texte.
- Choix des variantes.
- Apparat critique et notes de l'auteur.

Dans cet article, nous nous intéressons surtout à l'élaboration de *stemma codicum*. Cela consiste à trier les différentes versions du texte afin d'établir une sorte d'arbre de filiation des manuscrits du corpus pour savoir lequel a été copié sur l'autre et détecter les chaînons manquants. Il permet alors d'essayer de reconstituer le manuscrit original avec fidélité. L'analyse des différents manuscrits en vue d'une édition critique est un travail colossal lorsque le corpus est important. Ce travail peut être allégé par l'automatisation de certaines de ses parties ou par la production automatique de documents intermédiaires. Cet article décrit une famille de méthodes visant à faciliter la production du *stemma codicum*.

Les méthodes présentées ici empruntent largement à la phylogénétique. Buneman [1971(a)] et Griffith [1969] ont déjà fait remarqué la similitude des préoccupations

[2]La codicologie étudie le manuscrit comme objet matériel afin de mieux comprendre l'histoire du texte (ou des textes) qui est parvenu jusqu'à nous. C'est ainsi qu'elle étudie les techniques de fabrication et les divers accidents qui ont pu affecter ces ouvrages comme l'interversion de cahiers, la numérotation des folios, les accidents destructifs, les types de support...

[3]La paléographie a pour objet les écritures anciennes, leur déchiffrement, leur datation...

de l'édition critique et de la phylogénétique, celle-ci s'intéressant à la filiation entre les espèces vivantes, celle-là entre les manuscrits. Les méthodes de construction de l'arbre phylogénétique qui représente l'évolution des espèces peuvent être adaptées aux corpus de manuscrits, pour obtenir un arbre de la filiation.

FIGURE 1. Photo prise à Pune d'un fragment d'un manuscrit de la *kāśikāvṛtti*

Le plan de l'article est le suivant. Le Chapitre 2 est un rapide aperçu du travail de l'éditeur critique et plus particulièrement des méthodes utilisées pour dresser le *stemma*. Le Chapitre 3 s'intéresse aux manuscrits, aux particularités du sanskrit et aux différentes opérations préparatoires menées sur le corpus. Le Chapitre 4 présente plusieurs familles de mesures de similarité entre les manuscrits basées sur l'alignement et sur la compression des données. Le Chapitre 5 décrit les algorithmes d'inférence d'arbres utilisés en phylogénie, en particulier, le problème de la recherche de la racine. Finalement, le Chapitre 6 est consacré aux diverses expérimentations réalisées sur plusieurs corpus. La conclusion envisage des perspectives.

3. MÉTHODES PHILOLOGIQUES D'ÉTABLISSEMENT DU STEMMA CODICUM

Un texte qui a été copié des centaines de fois constitue ce que l'on appelle une *tradition textuelle* et tous les exemplaires qui nous sont parvenus sont appelés les *témoins* du texte.

Les témoins sont généralement différents. L'auteur a pu écrire différentes versions de son texte, les copistes ont fait des erreurs (oubli de mot, saut de ligne, amélioration...) et les évolutions du temps et de l'espace (trous dans le papier, paragraphe illisible, évolution de la langue...) multiplient les dissemblances entre témoins.

L'éditeur critique doit à, partir de ce corpus, reconstituer au mieux le manuscrit original encore appelé *archétype*[4]. Il réalise ce que l'on appelle le *texte critique*, c'est-à-dire, une version du texte la plus proche possible de l'archétype, en argumentant ses choix quand ils sont discutables.

Pour construire le texte critique à partir des témoins, on part du constat suivant : toutes les copies qui contiennent, aux mêmes endroits, les mêmes fautes, ont été faites les unes sur les autres et donc dérivent toutes d'une copie où ces fautes existaient. Pour classer les témoins, on recourt donc à la méthode de la comparaison

[4]L'archétype est, pour les éditeurs, le manuscrit d'où dérive les autres textes.

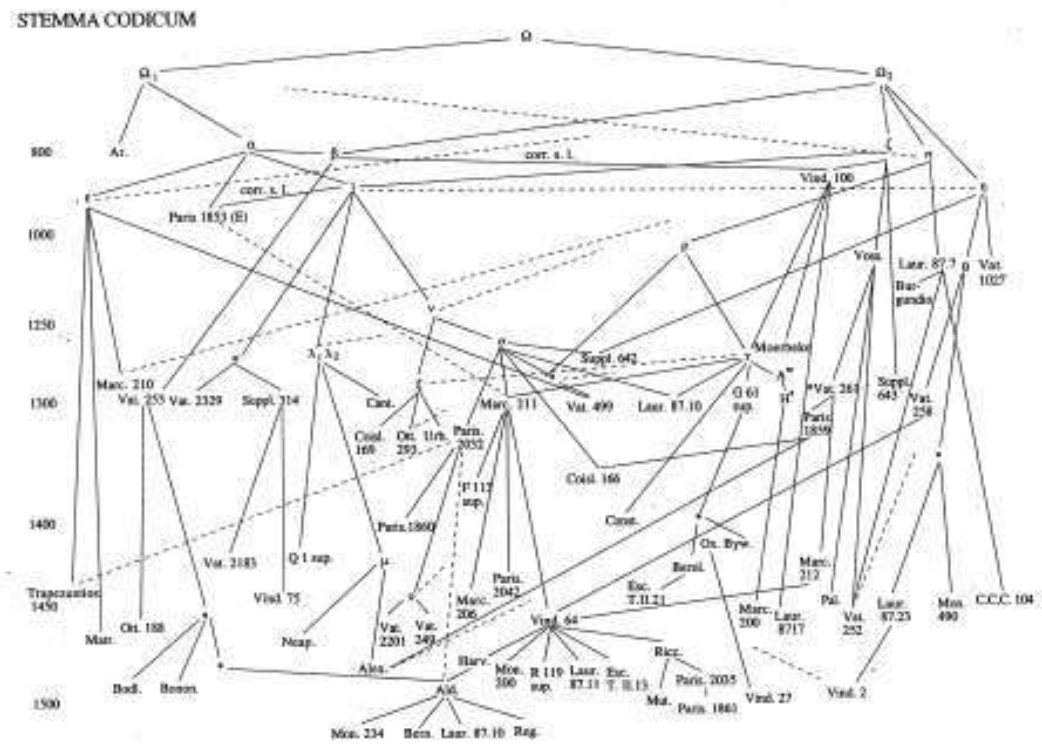


FIGURE 2. Stemma codicum de « De la génération et la corruption » d'Aristote établi par Rashed [2005]

des fautes appelées variantes qui, classées selon leur influence sur l'acte de copie, permettent de dresser un arbre généalogique des manuscrits. Cette méthode dite « lachmannienne » (cf. [Salemans, 2000]) présente l'avantage de préparer le travail de l'édition critique car, pour reconstituer le texte le plus proche de l'original, on évalue quelle variante convient le mieux.

Une autre méthode historique est celle de Don Quentin [1926]. Il s'attache à reconstituer l'enchaînement des manuscrits au moyen de comparaison trois par trois. C'est la recherche des intermédiaires qui permet de reconstituer le stemma. Pour que B soit intermédiaire entre A et C, il suffit que A et C s'accordent tour à tour avec B au niveau des variantes et surtout qu'ils ne s'accordent jamais contre lui. Il se propose alors de reconstituer des chaînes de trois manuscrits dont l'un est l'intermédiaire des deux autres puis d'assembler ces petites chaînes afin d'inférer l'arbre complet.

Prenons un exemple. Soient les trois phrases suivantes correspondants aux trois mêmes phrases de différents manuscrits copiés les uns sur les autres.

A = « Voici une phrase inventée pour l'exemple »

B = « Voici une phrase inventée pour cet exemple »

C = « Voici une phrase créée pour cet exemple »

La méthode de Don Quentin conduit à considérer que la phrase B est intermédiaire au sens de la copie entre A et C. En effet, le copiste de B a modifié « l' » en « cet » et le copiste de C a remplacé « inventée » par « créée ». B s'accorde avec A pour la variante « inventée » et s'accorde avec C pour la variante « créée ». C'est, en revanche, peu probable que C soit l'intermédiaire entre A et B, car le copiste de C a supprimé « inventée » qui est réintroduit par le copiste de B. Ici, A et B s'accordent sur la variante « inventée » qui n'est pas contenue dans C. C n'est donc pas le manuscrit intermédiaire entre A et B.

Plusieurs stemmata[5] différents peuvent être produits à partir d'une étude philologique selon le choix du point d'orientation, c'est-à-dire le choix du manuscrit ancêtre commun. Bédier [1928] l'a découvert lors de l'étude des textes du « Lai de l'Ombre ». Il a essayé de démontrer que des méthodes généalogiques, menant à beaucoup de résultats différents, sont sans valeur. Il a alors préconisé de s'en tenir aux variantes d'un témoin unique, celui qu'il a jugé le meilleur.

Finalement, on constate que lorsque le nombre de témoins est important ($\geqslant$100), les éditeurs ne construisent que très rarement le stemma pour les deux raisons principales suivantes:

- La collation de plus de cent manuscrits est un travail trop important manuellement.
- Le résultat s'avère difficilement exploitable tellement le stemma est complexe (cf. Figure 3.).

L'éditeur décide alors de sélectionner un petit nombre de manuscrits prépondérants pour effectuer l'édition. La méthode de détermination des manuscrits conservés n'est pas très scientifique (cf. [Bevenot, 1961]). C'est par des approches externes[6] que ce dernier a sélectionné sa famille de manuscrits, mais on peut aussi ne retenir que les manuscrits qui contiennent toutes les *leçons* (autre mot pour variantes) et éliminer les doublons.

Après ce rapide panorama, on cerne mieux la difficulté du problème, et l'on imagine le nombre d'années de travail nécessaires pour la réalisation d'une édition critique d'une tradition textuelle de 150 manuscrits composée de milliers de variantes. L'utilisation de méthodes informatiques pour aider l'éditeur critique s'impose alors.

4. LES MANUSCRITS ET LEURS TRAITEMENTS

Le corpus concerné dans cette étude est constitué de manuscrits sanskrits relatifs à la Glose de Bénarès. C'est le commentaire antique le plus complet sur le traité grammatical de *Pāṇini* (5^{e} siècle av. Jésus-Christ). Il existe environ 150 manuscrits

[5]Pour le pluriel de stemma, on peut utiliser la forme grecque avec *stemmata*, la forme latinisée avec *stemmæ* ou la forme francisée avec *stemmas*.

[6]Les approches externes concernent plus l'aspect extérieur des manuscrits comme la datation, le type d'écriture ou la provenance... Elle est à opposer à la structure interne des manuscrits comme les variantes ou les traits grammaticaux...

de la Glose de Bénarès recensés en Inde et en Occident. Chaque manuscrit comporte environ 800 pages, 245000 lignes et 1,5 millions de caractères. Pour notre étude, trois chapitres ont été sélectionnés d'environ 20 pages chacun.

On commence par l'exposé de certaines caractéristiques du sanskrit, dont la prise en compte est indispensable pour l'objectif visé. On réalise ensuite la translittération des différents manuscrits puis la segmentation des mots d'un des manuscrits que l'on nomme *padapāṭha*.

4.1. CARACTÉRISTIQUES DU SANSKRIT

Le sanskrit est une langue indo-européenne. C'est notamment la langue de la transmission de la connaissance savante en Inde, que cette connaissance soit profane (grammaire, mathématiques, architecture...) ou religieuse. Longtemps de tradition orale, c'est tardivement que l'emploi d'une multitude d'écritures s'est généralisé. Chaque région de l'Inde utilise la graphie qui lui sert pour noter sa propre langue afin d'écrire les textes sanskrits. Ainsi l'on ne compte pas moins d'une dizaine de graphies différentes pour nos manuscrits comme la *denavāgarī*, la *bengālī* ou le *tegulŭ*.

La principale graphie, la *denavāgarī*, est un des alphabets ou syllabaires utilisés par le sanskrit, le hindi et plusieurs autres langues indiennes. Elle comporte 56 graphèmes. Il s'agit de lettres (consonnes et voyelles, plus signes de ponctuations) et non pas d'un système d'idéogrammes comme en Mandarin.

अ	आ	इ	ई	उ	ऊ	ऋ	ॠ	ऌ	ॡ
a	ā	i	ī	u	ū	ṛ	ṝ	ḷ	ḹ

FIGURE 3. Voyelles de l'alphabet *denavāgarī* et leur transcription

Contrairement au français où les lettres s'ajoutent simplement les unes aux autres, la syllabe sanscrite est le fruit d'une ligature, comme notre Æ.

क	का	कि	की	कु	कू	कृ	कॄ	कॢ
ka	kā	ki	kī	ku	kū	kṛ	kṝ	kḷ

FIGURE 4. Les ligatures de la consonne K avec voyelles et leur transcription

Le sanskrit est une langue qui peut s'écrire sans espace ni ponctuation entre les mots. L'absence de blanc peut rendre les textes ambigus voire, dans certain cas, difficilement compréhensibles.

Ajoutons que les initiales et finales de chaque mot influent les unes sur les autres, apportant de nombreuses modifications ; d'où la première difficulté en lecture de reconnaître chaque terme et de restituer sa forme originale.

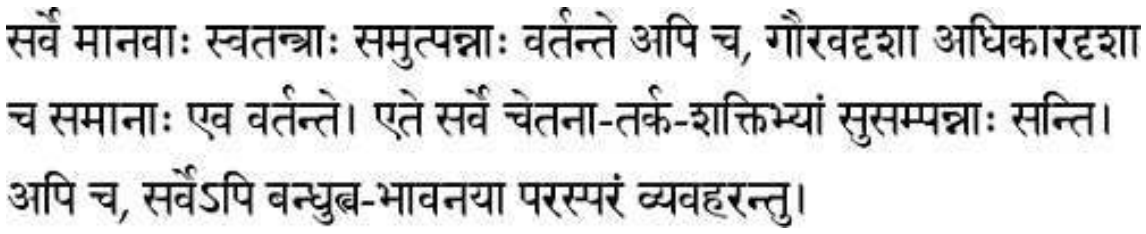


FIGURE 5. Exemple de sanskrit

Les *sandhi* (jonctions) sont une difficulté supplémentaire dans la reconnaissance des mots. Ils désignent des opérations phonétiques qui s'appliquent à la jonction de morphèmes à l'intérieur d'un mot où, à la jonction de mots dans la phrase. Quand ils s'appliquent entre des mots, ils correspondent à des modifications de la terminaison d'un mot selon la nature de la première lettre du mot lui succédant.

Pour en finir avec les spécificités du sanskrit, des spécialistes comme Filliozat [1941], qui parle de l'orthographe « vicieuse mais traditionnelle des scribes », ont montré que les mots difficiles à comprendre sont ceux où les manuscrits proposent une autre variante. Preuve que, loin de recopier sans comprendre, le scribe a remplacé le mot qu'il ne comprenait pas...

Pour pouvoir réaliser des analyses textuelles sur ces manuscrits, nous leur avons fait subir certaines opérations que nous allons détailler.

4.2. LA TRANSLITTÉRATION

La translittération, ou transcription, est l'opération consistant à transcrire les graphèmes d'un alphabet dans les graphèmes d'un autre alphabet, de telle sorte qu'à un même graphème ou à une suite de graphèmes de l'écriture de départ corresponde toujours un même graphème ou suite de graphèmes du système d'écriture d'arrivée. Nous utilisons alors un alphabet romain augmenté de signes diacritiques[7], qui permet de transcrire la *denavāgarī* sous une forme facilement lisible et imprimable pour des européens (cf. Figures 3 et 4).

Afin de procéder à la transcription, nous nous sommes contentés de suivre les conventions de transcription de F. Velthuis [1991] en utilisant le logiciel TeX/LaTeX. Cette procédure présente l'avantage de rendre les données utilisables sur n'importe quelle plate-forme sans conversion.

4.3. LA SEGMENTATION DES MOTS ET LE PADAPĀṬHA

L'absence de caractère séparateur complique énormément la reconnaissance des mots dans les différents manuscrits translittérés. Afin de rendre possible l'identification

[7]Un signe diacritique est un signe placé sur, sous, dans, après, etc, un graphème (exemple : en français, l'accent circonflexe).

```
{\rm [psu-5]}
ke.saa.m "sabdaanaa.m? laukikaanaa.m vaidikaanaa.m ca
kathamanu"saasana.m?  prak.rtyaadivibhaagakalpanayaa
saamaanyavi"se.savataa lak.sa.nena
{\rm [psu-6]}
pratyaahaaraprakara.nam
atha kimartho var.naanaamupade"sa.h ?  pratyaahaaraartha.h
pratyaahaaro laaghavena "saastraprav.rttyartha.h
a i u .n
{\rm [psu-7]}
a i u ityanena krame.na var.naanupadi"syaante .nakaaramita.m
karoti pratyaahaaraartham tasya graha.na.m bhavatyekena ura.n
apara.h ityakaare.na hrasvamavar.na.m prayoge
sa.mv.rta.m diirghaplutayostu viv.rtatvam te.saa.m
saavar.nyaprasiddhyarthamakaara iha "saastre
viv.rta.h pratij~naayate tasya prayogaartham a a iti
"saastraante pratyaapatti.h kari.syate
```

FIGURE 6. Exemple de données translittérées

des mots sur lesquels peut porter la comparaison entre les manuscrits, nous allons utiliser un texte segmenté nommé *padapāṭha*.

Le *padapāṭha* est une version du manuscrit réalisée par l'éditeur qui indique les séparations entre les mots, les racines, les préfixes etc. Il va nous permettre de comparer les manuscrits entre eux, au niveau des mots, en utilisant le *padapāṭha* comme un dictionnaire pour notre analyse.

Seul, à l'heure actuelle, manque un *segmenteur automatique* des mots qui permettrait alors une comparaison directe des manuscrits les uns avec les autres.

5. ALIGNEMENT, DISTANCE

5.1. AVERTISSEMENT AUX LECTEURS ET DÉFINITIONS

Nous avons besoin dans la suite de l'article du mot *distance* dans le sens courant et dans le sens mathématique. Pour éviter les confusions, nous réserverons le mot distance pour l'usage courant, c'est-à-dire, comme une mesure de la différence entre des objets. Nous utiliserons le mot *métrique* pour la définition mathématique suivante :

DÉFINITION 1. *Étant donné un ensemble E, une* métrique *est une application d de* $E * E$ *dans* $\mathbb{R}^{+}$[8] *qui vérifie les conditions suivantes :*

d1) d *symétrique :* $\forall (x, y) \in E * E \qquad d(x, y) = d(y, x)$ *(symétrie)*

d2) $d(x, y) = 0 \Leftrightarrow, x = y$ *(séparation)*

d3) $\forall x, y, z \in E \quad d(x, x) \leq d(x, z) + d(z, y)$ *(inégalité triangulaire)*

[8] $\mathbb{R}^{+}$ est l'ensemble des réels positifs ou nuls.

Si, à la place de la condition de séparation d2, l'application d vérifie la condition d4 suivante, on dit que c'est un écart.

d4) $\forall x \in E \qquad d(x,x) = 0$

L'application qui satisfait d1 et d4 est appelée dissimilarité. *Si, à la place de l'inégalité triangulaire d3, l'application vérifie la condition d5 suivante, on dit que c'est une* distance ultramétrique.

d5) $\forall x, y, z \in E \quad d(x,y) \leq max(d(x,z), d(z,y))$

5.2. INTRODUCTION À L'ALIGNEMENT ET AUX PROBLÈMES DE « GRANULARITÉ »

L'alignement permet la mise en correspondance de parties semblables des différentes versions d'un même texte. Pour cela, il convient de décider de la granularité de l'alignement, c'est-à-dire de la nature des parties ou segments à mettre en correspondance.

L'absence de segmentation explicite et systématique dans les manuscrits sanskrits représente une difficulté supplémentaire. La segmentation peut être effectuée à plusieurs niveaux : les caractères, les syllabes, les mots, les lemmes, les phrases, les paragraphes, etc. (cf. Figure 7). Par expérience, plus la segmentation est riche en sens (des lemmes plutôt que des mots), plus les informations sont pertinentes.

Dans un premier temps, nous nous contenterons de deux niveaux de base : les caractères et les mots à cause des difficultés liées au sanskrit.

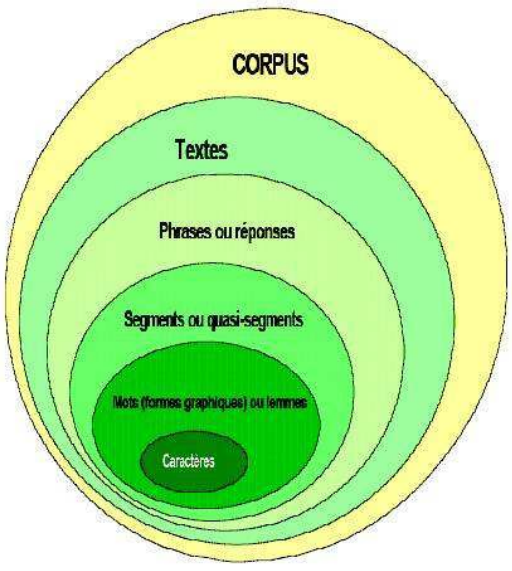


FIGURE 7. Différents niveaux de segmentation des textes

_	e	x	e	m	p	l	e
s	_	i	_	m	p	l	e

Table 1. Exemple d'alignement

5.3. ALIGNEMENT

Soit A un alphabet fini. Si t est une chaîne non vide sur A, on note $t[i]$ son $i^{\text{ième}}$ caractère. On suppose par ailleurs un caractère distingué, noté « $-$ », n'appartenant pas à A, qu'on appellera « blanc ».

DÉFINITION 2. *Soient deux chaînes $p1$ et $p2$ sur l'alphabet A. Un alignement de $p1$ et $p2$, pris dans cet ordre, est un couple de chaînes $(P1, P2)$ sur l'alphabet $A \cup \{-\}$ tel que :*

- *$P1$ et $P2$ sont de même longueur ;*
- *$P1$ est obtenu en insérant des blancs dans $p1$, autrement dit, on obtient $p1$ en effaçant toutes les occurrences du blanc dans $P1$;*
- *$P2$ est obtenu en insérant des blancs dans $p2$, autrement dit, on obtient $p2$ en effaçant toutes les occurrences du blanc dans $P2$;*
- *$\forall i$, non$((P1[i] = \text{« } - \text{ »})$ et $(P2[i] = \text{« } - \text{ »}))$, autrement dit un blanc ne se trouve jamais à la même position dans $P1$ et dans $P2$.*

Pour la suite, il est commode d'écrire les chaînes $P1$ et $P2$ d'un alignement l'une en dessous de l'autre. La Table 1 montre un alignement des chaînes *exemple* et *simple*. Cette manière d'écrire les alignements les fait apparaître comme des mots sur un alphabet dont les caractères sont des couples de caractères. On appellera *appariemen*t ces couples. Ainsi, l'alignement de la Table 1 peut-il être considéré comme un mot dont le premier caractère est l'appariement $(-, s)$, le deuxième l'appariement $(e, -)$, le troisième l'appariement (x, i) et ainsi de suite.

Tout alignement des phrases $p1$ et $p2$ peut s'interpréter comme une transformation agissant sur $p1$ et produisant $p2$. Pour ce faire, on associe à chaque appariement une opération d'édition élémentaire (substitution d'un caractère à un autre, insertion ou suppression d'un caractère ou identité). Dès lors, la transformation associée à un alignement est la séquence des opérations associées à chacun de ses appariements. Ainsi, à partir de l'exemple de la Table 1, le premier appariement $(-, x)$ s'interprète comme l'insertion dans `exemple` du caractère `s` ; on obtient `sexemple`. Le deuxième appariement $(e, -)$ s'interprète comme la suppression du caractère `e` ; on obtient `sxemple`. Le troisième appariement (x, i) permet de remplacer la caractère `x` en `i` ; on obtient `siemple`. Le quatrième appariement $(e, -)$ supprime le caractère `e` ; on obtient `simple`. Les appariements qui suivent sont tous de la forme (x, x) : ils laissent inchangée la chaîne source, ce sont des identités. Finalement, la transformation associée à l'alignement aboutit à la chaîne `simple`.

Le nombre d'alignements possibles entre deux chaînes de longueur l1 et l2 est donné par la fonction de Waterman [1995] :

$$F(l1, l2) = \sum_{k=0}^{\min(l1,l2)} \frac{(l1 + l2 - k)!}{k!\ (l1 - k)!\ (l2 - k)!}$$

Il peut devenir très grand avec la longueur des chaînes. Il vaut 681 pour l1 = 4 et l2 = 5 et 265729 pour l1 = l2 = 8. Cette combinatoire doit, à l'évidence, être réduite. À quelques exceptions près, un sanskritiste produirait un seul alignement, voire deux. Comment simuler son travail ? Comment sélectionner parmi tous les alignements possibles celui (ceux) du sanskritiste ? La tentative formelle décrite ici s'inspire directement des distances d'édition (cf. [Levenshtein, 1966]). On commence par présenter la méthode retenue (et mise en œuvre) pour ensuite discuter de son adéquation. Elle se décompose en trois moments :

- premier moment : on considère une application qui associe à chaque appariement un entier ≥ 0, son poids ;
- deuxième moment : on associe, à chaque alignement, une quantité appelée son poids qui est égale à la somme des poids de ses appariements ;
- troisième moment : on retient comme critère de sélection des alignements celui qui consiste à considérer l'ensemble de ceux de poids minimal.

Il a déjà été dit qu'un alignement pouvait s'interpréter comme une transformation d'édition. Le poids d'un alignement est donc aussi le coût de la transformation d'édition associée. Les alignements sélectionnés par le critère décrit ci-dessus sont donc ceux qui correspondent aux transformations d'édition les moins coûteuses.

Le modèle d'alignement le plus simple est celui introduit en 1965 par Levenshtein. Il définit le poids d'un alignement comme la quantité $l - i$, l étant la longueur de l'alignement et i le nombre de ses appariements de la forme (x,x), autrement dit la même lettre en positions haute et basse. Ce poids est encore égal au nombre d'appariements de la forme (x, y), $x \neq y$, donc aussi au nombre de transformations d'édition élémentaire (insertion, suppression, substitution) que l'opération d'édition associée à l'alignement utilise. Finalement, il se calcule en affectant aux appariements (x, y), $x \neq y$, le poids 1 et aux appariements (x, x) le poids 0. Étant données deux chaînes, Levenshtein considère la quantité égale au minimum des poids de leurs alignements. Elle a les propriétés d'une métrique et peut être calculée à partir d'un algorithme de programmation dynamique (cf. [Bellman, 1957 ; Wagner et Fisher, 1974]).

Dans le cas qui nous occupe, les spécificités du sanskrit et de la translittération gênent l'utilisation des alignements classiques et on peut se poser quelques questions :

- Le nombre d'opérations élémentaires est-il satisfaisant ?

  Certains caractères sanskrits peuvent correspondre à plusieurs caractères latins du fait de la translittération. Un alignement au niveau des caractères latins ne correspond pas forcément à l'alignement au niveau des caractères sanskrits qui est préférable. Par conséquent, un alignement basé sur les caractères latins peut ne pas correspondre à un alignement pertinent des caractères sanskrits.

EXEMPLE. Soient les deux textes sanskrits translittérés en Velthuis $p1$ = `yamaan` et $p2$ = `yamin` à aligner. On obtient l'alignement (a) et la métrique entre $p1$ et $p2$ est 2. Or, dans $p1$, la sous-chaîne `aa` correspond à un seul caractère sanskrit आ. Dans la translittération IAST [9], ce caractère s'écrirait avec un seul caractère `ā` et $p1$ deviendrait $p1'$ = `yamān`. L'alignement de $p1'$ et de $p2$ est celui de la table (b) et la métrique entre $p1'$ et $p2$ est 1.

y	a	m	a	a	n
y	a	m	_	i	n

(a)

y	a	m	ā	n
y	a	m	i	n

(b)

On constate que l'alignement au niveau de la translittération IAST correspond mieux à celui que l'on obtient avec les caractères sanskrits ; la métrique induite reste identique. De plus, l'alignement obtenu par la translittération Velthuis n'est pas unique.

- Quel poids doit-on affecter aux opérations élémentaires pour que cela corresponde au meilleur alignement au niveau du sanskrit ?

  Considérons les deux mots suivants `"srii gurave` et `"srii ga.ne"saaya`. On peut les aligner des deux façons suivantes :

"s	r	ii		g	u	r	a	-	v	e	-	-	-	-
"s	r	ii		g	-	-	a	.n	-	e	"s	aa	y	a

(a)

"s	r	ii		g	-	-	u	r	a	v	e
"s	r	ii		g	a	.n	e	"s	aa	y	a

(b)

L'alignement (a) est celui dans lequel le plus grand nombre de symboles alignés a été conservé (6 bien alignés contre 8 erreurs), alors que l'alignement (b) est celui dont le nombre d'erreurs d'alignement est minimal (5 bien alignés contre 6 erreurs). Ces deux alignements sont optimaux selon le critère défini pour les construire mais pas identiques. De l'avis des sanskritistes, le deuxième alignement est le plus pertinent. Quels poids faut-il affecter aux opérations d'édition pour l'obtenir ?

- L'alignement doit essayer de conserver les mots dans la mesure du possible. Dans l'exemple suivant, le premier alignement « éclate» le mot *gurave* alors que l'on préférerait l'autre alignement.

[9] L'IAST est une autre norme de translittération du sanskrit établi en 1912 au Congrès des Orientalistes de Athènes.

g	u	-	-	-	-	-	r	-	-	-	-	-	-	-	-	a	v	e
g	-	u	d	i	p	a	r	s	v	a	n	a	h	a	y	a	-	-

(a)

g	u	r	a	v	e	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	g	u	d	i	p	a	r	s	v	a	n	a	h	a	y	a

(b)

En plus, comme le sanskrit n'a pas de séparateurs entre les mots, il faut impérativement utiliser le *padapāṭha* (cf. 4.3.) pour optimiser cette contrainte.

- Une autre difficulté provient des *sandhi* (cf. 4.1.). Si un des mots disparaît dans une version, il y a constitution éventuelle d'un nouveau *sandhi* dépendant des mots se trouvant alors en contact. La disparition d'un mot entraîne donc une série de modifications supplémentaires (outre sa propre suppression) sur le texte ce qui complique l'alignement.

- Enfin, les manuscrits lacunaires soulèvent un problème particulier. Si les paragraphes mis en parallèle n'ont pas sensiblement la même longueur, l'alignement peut se détériorer. Nous proposons donc de commencer par un alignement des phrases à l'intérieur de chaque paragraphe, ce qui permettra de supprimer de l'alignement les phrases considérées comme trop lacunaires. On aligne ensuite uniquement les phrases considérées comme « alignables ».

Voyons maintenant les méthodes proposées :

1^e méthode : alignement de corpus multilingues + distance d'édition

Comme nous l'avons évoqué précédemment, l'utilisation de l'alignement par la distance d'édition fonctionne très bien sur des petites parties de textes mais devient difficilement calculable et de mauvaise qualité sur des textes plus longs. Afin d'éviter cette difficulté, on procède, par une technique autre que celle de la distance d'édition, à un premier alignement des phrases à l'intérieur de chaque paragraphe, puis on utilise les distances d'éditions sur les phrases mises en parallèle.

Nous utilisons les techniques de l'alignement de corpus multilingues afin d'aligner les phrases des différents manuscrits. C'est la méthode statistique de Gale et Church [1991] qui a été utilisée. Ce type d'alignement utilise uniquement les longueurs des phrases et aucune information sur leur contenu lexical. Les auteurs partent de la constatation que la longueur des phrases du texte source est fortement corrélée avec celle des phrases du texte cible. Ils ont d'ailleurs montré que cette corrélation suivait une loi de probabilité normale centrée réduite.

Bien que simple, ce critère donne de très bons résultats lorsque les langues sont proches. C'est aussi le cas pour l'alignement des phrases de notre corpus. Il permet en particulier d'écarter les phrases non « alignables ».

La Figure 8 contient les résultats fournis par cet algorithme entre les trois premiers paragraphes de deux manuscrits. Chaque ligne du premier manuscrit est précédée du symbole >>, chaque ligne du second, du symbole <<. On observe que

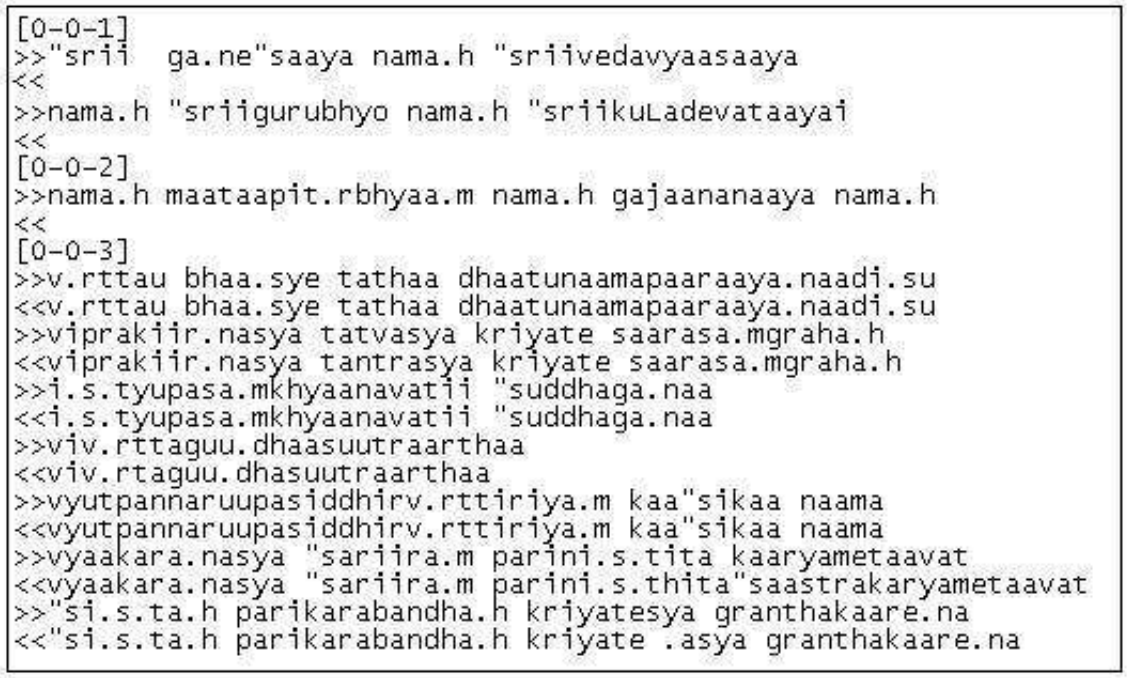


FIGURE 8. Résultats de l'algorithme de Gale et Church

les alignements calculés sont satisfaisants. On souligne de plus que l'absence des deux premiers paragraphes dans le second manuscrit, est détectée en ce sens que leurs phrases ne sont alignées à aucune phrase du second manuscrit.

On peut, par la suite, réaliser l'alignement des manuscrits au niveau des caractères pour chaque couple de phrases « alignables » par la métrique de Levenshtein [1966]. Une distance entre les manuscrits peut être construite en sommant les métriques obtenues sur l'ensemble des phrases « alignables ».

2^e MÉTHODE : COMPARAISON DES MANUSCRITS À L'AIDE DU *padapāṭha* + DISTANCE SUR LES MOTS

Une méthode intéressante de collation, actuellement développée par Csernel et Bertrand [2005], fournit pour tout manuscrit un alignement avec le *padapāṭha* qui respecte à la fois le découpage en mots de ce dernier et l'ordre de ses mots. Cette collation permet d'associer à chaque manuscrit une suite de valeurs binaires 0 ou 1, dont la longueur est celle du *padapāṭha* (considéré comme une suite de mots) et qui contient d'autant plus de 1 que le manuscrit concerné « ressemble » au *padapāṭha*. On se propose alors d'utiliser la métrique de Hamming [1950] entre les suites binaires associés aux manuscrits pour évaluer une distance entre les manuscrits.

Formellement, on considère un alphabet C de caractères et un ensemble fini D (dictionnaire) de mots sur C. On désigna alors par P (P correspond au *padapāṭha*) une suite finie d'éléments de D de longueur p. On note $P[i]$ son $i^{\text{ième}}$ mot. On considère, de plus, une suite finie M d'éléments de C de longueur m (M correspond à un manuscrit). On désigne par $M[j]$ son $j^{\text{ième}}$ caractère et par $M[j_1, j_2]$, $j_2 \geq j_1$, le mot formé par les caractères $M[j_1]M[j_1+1]M[j_1+2]...M[j_2]$.

On appelle *mises en correspondance* de M avec P, toute relation ternaire R de $\mathbb{N}^3$ telle que :

1. $R(i, j_1, j_2) \Rightarrow (1 \leq i \leq p)$ et $(1 \leq j_1 \leq j_2 \leq m)$
2. $R(i, j_1, j_2) \Rightarrow P[i] = M[j_1, j_2]$
3. $R(i, j_1, j_2)$ et $R(i, j'_1, j'_2) \Rightarrow (j_1 = j'_1)$ et $(j_2 = j'_2)$
4. $R(i, j_1, j_2)$ et $R(i', j'_1, j'_2)$ et $(i < i') \Rightarrow (j_2 < j'_1)$

La relation R traduit le fait que le mot $M[j_1, j_2]$ formé par les $j_2 - j_1$ caractères à partir du caractère j_1 dans M est égale au mot en position i dans P (propriété 2). Elle est fonctionnelle en i (propriété 3) et les mots construits à partir des caractères de M sont disjoints et respectent l'ordre des mots de P (propriété 4).

On appelle *masque binaire* de R la suite binaire finie S de longueur p telle que

$$S[i] = \begin{cases} 1 & \text{si } \exists (j_1, j_2)\, (j_2 > j_1)\, \text{tq}\, R(i, j_1, j_2) \\ 0 & \text{sinon} \end{cases}$$

Pour un manuscrit donné, il existe plusieurs mises en correspondance avec le texte de référence P, ne serait-ce que la mise en correspondance triviale dans laquelle la relation R est vide. L'intérêt de l'algorithme de Csernel et Bertrand est de sélectionner, parmi toutes les mises en correspondance possibles, une seule d'entre elles, considérée comme la meilleure. Il permet donc d'associer à chaque manuscrit un seul masque binaire. Ce masque est un codage avec perte du manuscrit initial, puisqu'on ne peut pas reconstruire ce dernier à partir de son masque et du *padapāṭha*.

Si $S1$ est le masque associé à un manuscrit $M1$ et $S2$ celui associé au manuscrit $M2$, la métrique de Hamming [1950] entre les masques $S1$ et $S2$ (on rappelle qu'elle est égale au nombre d'indices i pour lesquels $S1[i] \neq S2[i]$) est une métrique entre les masques. Ce n'en est pas une entre les manuscrits. On peut cependant tenter de l'utiliser pour apprécier leur éloignement. Si les manuscrits sont trop éloignés du document de référence P, les résultats risquent d'être médiocres. Ce n'est pas le cas dans notre application, puisque les manuscrits sont proches les uns des autres et le *padapāṭha* convenablement choisi.

EXEMPLE. Imaginons que le *padapāṭha* et deux manuscrits *pad*, $ms1$ et $ms2$ soient réduits à une phrase :

pad = *des équipes en Inde récupèrent les manuscrits*
ms_1 = *despetiteséquipesrécupèrentlesmanuscrits*
ms_2 = *deséquipesrécupèrentlaGlosedeBénarès*

Appliqué aux documents ms_1 et ms_1, l'algorithme de Csernel et Bertrand conduit aux mises en correspondance suivantes avec le *padapāṭha* :

$pad = \underline{des}$———-$\underline{équipes}\,\underline{en}\,\underline{Inde}\,\underline{récupèrent}\,\underline{les}\,\underline{manuscrits}$
$ms_1 = \underline{des}\,\underline{petites}\,\underline{équipes}$———-$\underline{récupèrent}\,\underline{les}\,\underline{manuscrits}$

$pad = \underline{des}\,\underline{équipes}\,\underline{en}\,\underline{Inde}\,\underline{récupèrent}\,\underline{les}\,\underline{manuscrits}$
$ms_1 = \underline{des}\,\underline{équipes}$———-$\underline{récupèrent}\,laGlosedeBénarès$

D'où les masques binaires des manuscrits 1 et 2, dessinés verticalement dans la Table 2. La métrique de Hamming entre les deux masques est 2.

Mot du *padapāṭha*	Manuscrit 1	Manuscrit 2
des	1	1
équipes	1	1
en	0	0
Inde	0	0
récupèrent	1	1
les	1	0
manuscrits	1	0

Table 2. Table binaire de présence ou d'absence des mots au cours de l'alignement

La distance de Hamming entre les masques binaires est alors égale à 2 ce qui correspond aux deux mots du *padapāṭha*, `les` et `manuscrits` qui appartiennent à l'un des manuscrits et pas à l'autre.

REMARQUE. La distance, qui est construite, est bien une métrique entre les deux masques binaires, mais elle n'est pas une métrique entre les deux manuscrits puisque le masque binaire est construit par un codage avec perte. Par cette méthode, la comparaison de deux manuscrits très éloignés du *padapāṭha* risque de ne pas être pertinente. En effet, la zone de comparaison (intersection des deux manuscrits avec le *padapāṭha* sur la Figure 9) doit être la plus grande possible pour que la distance entre les masques binaires soit la plus proche possible d'une distance entre les manuscrits. Dans notre cas, les manuscrits du corpus étant très proches les uns des autres et le *padapāṭha* convenablement choisi, les résultats obtenus restent cohérents.

5.4. DISTANCE DE COMPRESSION

Une autre sorte de distance peut être envisagée : la distance de compression. Contrairement aux distances précédentes, elle n'est pas basée sur des alignements ce qui s'éloigne de la démarche de l'éditeur. Deux raisons expliquent que nous ayons naturellement tendance à utiliser des distances issues de l'alignement :

- la première vient du fait que la collation des manuscrits demande un alignement de ceux-ci pour comparer les variantes ;
- la deuxième est issue de la comparaison avec les méthodes phylogénétiques qui, comme nous le verrons par la suite, utilisent, elles aussi, les méthodes d'alignement.

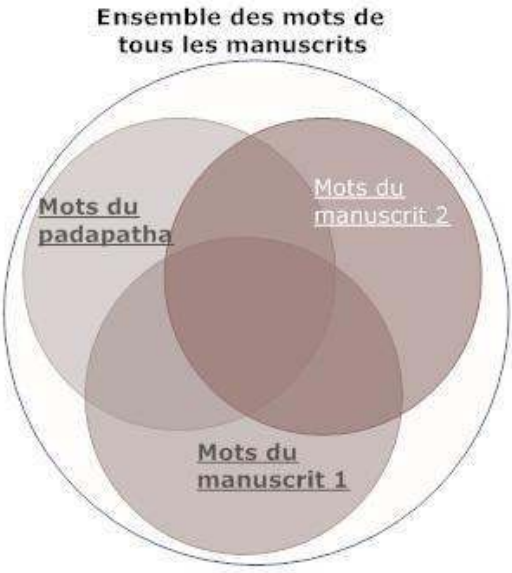


FIGURE 9. Schéma explicatif

Dans un article, Cilibrasi et Vitányi [2005] évoquent la possibilité d'appliquer les algorithmes de compression pour évaluer la distance normale de compression (NCD) entre divers types de documents, en particulier des textes. La distance est calculée par des algorithmes de compression qui fournissent une bonne évaluation de la quantité d'information contenue dans les données à classer.

Nous savons tous intuitivement ce qu'est une information, mais il est difficile d'en donner une définition précise. On sait dire que telle phrase contient plus d'information que telle autre. Par exemple, la phrase « Une équipe en Inde est chargée de récupérer les manuscrits » est plus informative que la phrase « Une équipe est chargée de récupérer les manuscrits ». La mesure de la quantité d'information d'un message peut s'effectuer par différentes méthodes théoriques :

- quantité statistique d'information au sens de Shannon [1948]
- théorie de la complexité de Kolmogorov… [1965]

Pour obtenir la quantité d'information d'une chaîne de caractères, l'approche de Shannon ne se focalise pas sur la chaîne de caractères elle-même, mais considère uniquement les probabilités d'obtenir cette chaîne de caractères parmi un ensemble de chaînes de caractères possibles. Shannon a démontré que cette quantité a une valeur maximale qu'il appelle entropie. L'approche proposée par la théorie de la complexité de Kolmogorov définit la quantité d'information d'une chaîne de caractères par la longueur du plus petit programme qui produit cette chaîne. Même si les deux démarches semblent éloignées, elles sont en fait compatibles et l'entropie de Shannon correspond à la complexité de Kolmogorov moyenne (cf. [Grünwald et Vitányi, 2004]). La complexité de Kolmogorov n'est pas calculable (cf. [Grünwald et Vitányi, 2004]), mais peut être approchée par des algorithmes de compression sans perte.

La compression est utilisée pour réduire la taille physique d'une suite d'octets. Il existe deux sortes de compression :

- La compression sans perte ou conservatrice (comme zip) consiste à déterminer un codage optimum permettant de stocker les données de telle manière que celles-ci soient récupérables dans leur intégralité. Il existe de nombreuses méthodes de compression sans perte permettant de d'approcher la complexité de Kolmogorov.
- La compression avec perte ou non conservative (comme jpg ou mp3) admet une différence ou distorsion entre les données originales et celles décodées. Elle accepte une légère perte de d'information afin de faciliter la compression. La théorie de l'information établit qu'il est possible d'introduire des distorsions ou pertes afin d'obtenir une compression inférieure à l'entropie.

Pour la suite, nous nous intéressons surtout à la compression conservative. Trois algorithmes de compression ont été envisagés.

méthode RLE (Run Length Encoding)

Cette technique consiste à repérer et à éliminer la redondance des données. C'est une méthode utilisée par de nombreux formats d'images (BMP, PCX, TIF). Elle est basée sur la répétition de bits consécutifs. Une première valeur (codée sur un octet) donne le nombre de répétitions, une seconde la valeur à répéter (codée elle aussi sur un octet).

exemple. La phrase suivante « yyyyyeeeaaahhhh » donne une fois codée : « 5y3e3a4h ». Par contre, la phrase « bonjour » donne « 1b1o1o1j1o1u1r » qui n'est pas du tout intéressante car elle rallonge le codage initial. Ce codage est plus adapté au codage des images où les couleurs qui ne varient pas sur plusieurs pixels qu'au format texte.

le codage Huffmann [1952]

Cette méthode de codage est de plus en plus utilisée à travers l'algorithme de Burrows et Wheeler [1994] qui donne la compression BZIP2. Pour coder les caractères de l'alphabet d'un fichier texte, nous associons à chacun d'entre eux un mot binaire. Dans une représentation habituelle, la longueur du mot binaire est la même pour tous les caractères, un octet par exemple. Le principe du codage de Huffman est d'associer à chaque caractère un code binaire qui est d'autant plus court que le caractère apparaît souvent dans le texte. Un code de Huffman consiste à établir un arbre binaire dont les feuilles sont étiquetées par les caractères du document. Le mot binaire associé à chaque caractère est le mot du chemin qui lie la racine à la feuille dont ce caractère est l'étiquette. Nous représentons alors une chaîne de caractères par la concaténation des mots binaires correspondant à chacun de ses caractères.

exemple. Le mot « trotter » contient trois **t** deux **r** un **o** et un **e**. Le **t** étant le plus présent, il est codé le plus simplement par le **1**... L'algorithme de Huffman construit l'arbre binaire de la Figure 10 qui associe aux caractères les plus fréquents le codage

le plus court. Dès lors, « trotter » s'écrit 1-01-000-1-1-001-01, qui n'utilise que 18 positions binaires alors que le codage usuel en utilise $7 * 8 = 56$.

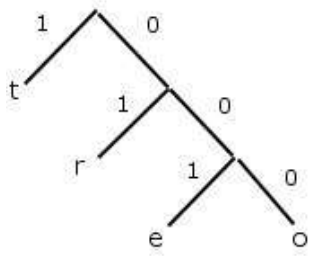


FIGURE 10. Arbre binaire associé aux caractères

LES ALGORITHMES À DICTIONNAIRE

Les algorithmes à dictionnaire, consistent à remplacer des séquences par un code plus court qui est l'indice de la séquence dans un dictionnaire construit au fur et à mesure de la compression du fichier. La compression LZ de Lempel et Ziv [1977] est utilisée par les formats GIF et TIFF et dans les compresseurs ZIP. Le gain de compression est énormément élevé par rapport à un algorithme traditionnel. L'avantage de la compression LZ est aussi qu'elle ne nécessite qu'une lecture du fichier (contrairement au codage de Huffman).

EXEMPLE. Soit le mot « bonbon ». Tous les caractères sont déjà dans le dictionnaire avec leur code de 1 à 26 pour les lettres

Caractère	Chaîne	Est-il dans le dictionnaire ?	Ajouter dans le dictionnaire	Code
b		oui		2
o	bo	non	bo	27
n	on	non	on	28
b	nb	non	nb	29
o	bo	oui		
n	bon	non	bon	30

Table 3. Codage par algorithme à dictionnaire du mot *bonbon*

Les autres chaînes de caractères sont codées dans le Tableau 3, et le codage de « bonbon » est 2 15 14 30.

Il apparaît, dans le cas d'élaboration d'une distance entre des manuscrits copiés les uns sur les autres, plus pertinent d'utiliser les algorithmes à dictionnaire, comme ZIP car ils sont mieux adaptés pour rendre compte de la différence de quantité d'information (chaine de caractères commune) contenue dans les manuscrits.

Une fois l'algorithme de compression déterminé, il reste à définir la distance. Notons $c(m1)$ la taille (en octets) du fichier $m1$ comprimé. Nous partons alors de l'axiome suivant : la compression d'un fichier est inférieure à la compression de la

concaténation de ce fichier et d'un autre. Soit $m1$ et $m2$ deux manuscrits, alors $c(m1) \leq c(m1 + m2)$ et cela correspond au fait que m2 augmente l'information contenue dans $m1$. Nous pouvons donc dire que la quantité $c(m1+m2)-c(m1)$ est la quantité d'information de m2 qui n'est pas dans m1. Nous proposons la quantité suivante :

$$D(m1, m2) = c(m1 + m2) + c(m2 + m1) - c(m1) - c(m2)$$

Pour être une métrique, $D(x, y)$ doit vérifier 3 axiomes de la Définition 1.

$D(x, y)$ vérifie l'axiome de symétrie. Les deux autres axiomes ne sont pas vérifiés. Cependant Bennett *et al.* [1998] s'approchent de la notion de métrique par la définition d'une *métrique admissible* que vérifie la distance de compression théorique (c'est-à-dire sans approximation par la compression). Dans la pratique, on peut dire que l'on a une quasi-dissimilarité car les valeurs de $D(m1, m1)$ sont quasi-nulles et donc très proches de l'axiome $d4$.

6. LES ARBRES

6.1. LA PHYLOGÉNIE

La phylogénie peut être considérée comme la construction de l'histoire évolutive d'un ensemble d'espèces. La plupart des méthodes phylogénétiques représentent les relations qui existent entre les espèces sous forme d'un arbre phylogénétique binaire[10], résolu[8] pour les biologistes et bifide[8] pour les éditeurs.

Un arbre phylogénétique est une représentation graphique de la phylogenèse d'un groupe d'espèces. C'est-à-dire : les feuilles représentent les espèces et les branches définissent les relations de parenté entre elles. Les sommets intérieurs autres que les feuilles représentent des ancêtres hypothétiques.

Il est intéressant de comparer cette théorie de l'évolution avec la filiation de manuscrits comme l'ont suggéré Buneman [1971(a)] et Griffith [1969]. En effet, les séquences génétiques de différentes espèces sont comparées et, selon leur ressemblance ou dissemblance, nous déterminons alors un arbre de l'évolution. Une situation similaire se retrouve si nous comparons les écrits de différents témoins les uns avec les autres et, peut être, pouvons-nous alors déterminer l'arbre de filiation des différents manuscrits ?

Cependant, une différence est à noter entre notre problème et celui de la théorie de l'évolution. Dans cette dernière, les espèces dont nous comparons les séquences d'ADN sont toutes à l'instant t de l'évolution donc, dans l'arbre, elles sont représentées par des feuilles. Dans la filiation de manuscrit, rien n'empêche de trouver un manuscrit qui soit « l'ancêtre commun » ou une copie intermédiaire d'autres manuscrits ; dans ce cas, le manuscrit est associé à la racine ou à un nœud de l'arbre (cf. Figure 11).

[10]Binaire, résolu et bifide sont des synonymes qui désignent un arbre avec racine dans lequel chaque nœud a au plus deux arêtes.

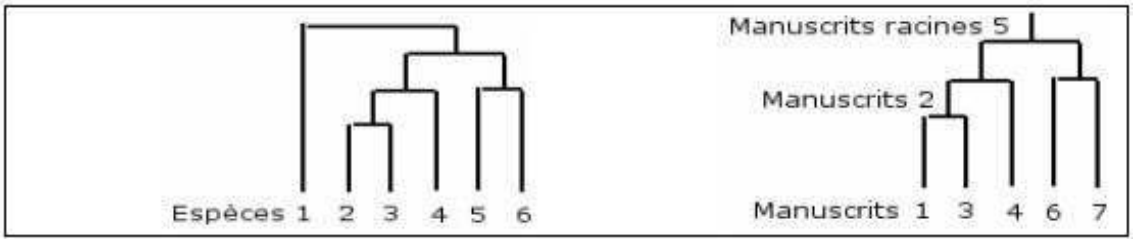


FIGURE 11. Différence entre phylogénie et filiation

Il existe différentes méthodes permettant de construire des arbres phylogénétiques :

- La méthode du maximum de parcimonie (*maximum parsimony*). Elle sélectionne l'arbre nécessitant le minimum de changements évolutifs.
- Les méthodes probabilistes en particulier la méthode du maximum de vraisemblance (*maximum likelihood*). Elles sont appliquées à un ensemble de caractères pour chacun desquels, une probabilité de transition entre les divers états est définie. À partir de là, la phylogénie est représentée par l'arbre dont la vraisemblance calculée à partir des probabilités est maximale.
- Les méthodes basées sur les distances (*pairwise distances*) se proposent de reconstruire des arbres, en partant des ressemblances observées entre chaque paire d'unités évolutives.

Au vu du nombre important de manuscrits, nous privilégions une méthode basée sur les distances. Il peut malgré tout être intéressant d'utiliser les autres méthodes pour confirmer les hypothèses faites avec la méthode des distances sur une partie de l'arbre (sous-arbre).

6.2. RAPPEL SUR LES ARBRES

Au niveau des notations, nous reprenons en partie celles décrites dans Barthélemy et Guénoche [1988].

DÉFINITION 3. *Un* graphe *G est un couple $G = (V, E)$ dans lequel V est un ensemble fini, l'ensemble des* sommets*, et E un ensemble de sous-ensembles $\{x, y\}$ de V à deux éléments, l'ensemble des* arêtes *(les arêtes sont des paires= ensembles à deux éléments, pas des couples). Pour simplifier, une arête de G est notée vv'. Le* degré *d'un sommet v est le nombre d'arêtes contenant v. Une* feuille *est un sommet de degré 1 et les autres sommets sont appelés des* nœuds*.*

Dans un graphe G, une *chaîne* P entre deux sommets v et v' est une suite finie de longueur ≥ 2 de sommets $(v_i; 0 \leq i \leq n)$, telle que $v_0 = v$ et $v_n = v'$ et pour tout i, $0 \leq i < n$, la paire v_i, v_{i+1} est une arête. Un chemin est une chaîne dont les arêtes sont toutes distinctes. Ainsi tout chemin pourra être identifié à l'ensemble de ses arêtes. La longueur d'un chemin est égale au nombre de ses arêtes. Si $v = v'$ dans

Nombre de feuilles	**Nb d'arbres non plantés**	**Nb d'arbres plantés**
2	1	1
4	3	15
5	15	105
8	10 395	135 135
10	34 459 425	$2.13\ 10^{15}$
15	$2.13\ 10^{15}$	$8.00\ 10^{21}$
n	$\prod_{i=2\cdots n}(2*i-5)$	$\prod_{i=2\cdots n}(2*i-3)$

Table 4. Nombre d'arbres possibles en fonction du nombre de feuilles

un chemin P alors c'est un *cycle*. Le graphe G est *connexe* s'il existe un chemin entre toute paire de sommets distincts de G.

DÉFINITION 4. *Un graphe G est un* arbre *s'il est connexe et n'a pas de cycles. Tout arbre T qui ne possède qu'un seul nœud est appelé une* étoile. *Un* arbre planté *ou* enraciné *est un couple (T, r) formé d'un arbre T et d'un sommet r appelé* racine. *Un arbre $T = (V, E)$ a exactement $|V|$-1 arêtes et il y a un chemin unique, noté $T(vv')$, entre deux sommets distincts v et v'.*

Le nombre d'arbres possibles pour n feuilles augmente rapidement (cf. Table 4) ; ce qui pose des problèmes de complexité informatique.

DÉFINITION 5. *Un* arbre valué *est un couple (T, L), où T est un arbre (V, E) et L une application qui associe, à chaque arrête, un réel positif appelé sa longueur. L'application L définit canoniquement une métrique d_L sur l'ensemble des sommets de l'arbre égale à la somme des longueurs des arêtes de l'unique chemin entre deux sommets.*

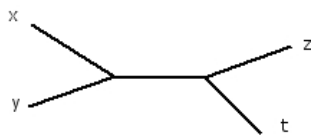
DÉFINITION 6. *Soit X un ensemble fini. Un* X-arbre *est un couple(T, f) formé d'un arbre $T = (V, E)$ et d'une fonction f de X dans V telle que pour tout $v \in V \backslash f(X)$, degré(v) = 3. La fonction f est* l'étiquetage *du X-arbre, les sommets dans $f(X)$ sont appelés* sommets réels,*les sommets de $V \backslash f(X)$ sont appelés* sommets latents.

Si n est le cardinal de X, un X-arbre a au plus 2n-2 sommets et donc 2n-3 arêtes. Pour tout triplet u, v, w de sommets distincts d'un arbre T, il y a un sommet unique $m(u, v, w)$ commun aux chemins uv, vw et uw appelé *médiane*.

DÉFINITION 7. *Une dissimilarité sur X est une* distance d'arbre ou arborée *si elle vérifie la* condition des quatre points *:*

$$\forall (x, y, z) \in X \qquad d(x, y) + d(z, t) \leq \sup\left(d(x, z) + d(y, t), d(x, t) + d(y, z)\right).$$

Graphiquement :



Le résultat central pour le propos de l'article est établi par Buneman [1971(b)] :

THÉORÈME 1. *Si d est une distance d'arbre, alors il existe un unique X-arbre valué (V, E, f, L) tel que $\forall (x, y) \in X, \quad d(x, y) = d_L(f(x), f(y))$ avec f injective (il n'y a pas d'étiquetage multiple).*

Ce théorème énonce comment une donnée numérique, la distance arborée sur X, est en bijection avec une donnée structurelle, l'X-arbre valué. Il est en parfaite adéquation avec le but visé ici, à savoir l'obtention d'un stemma codicum, donnée structurelle, fût-il approché, à partir des distances obtenues au chapitre précédent. Cela étant, rares sont les cas pratiques où la condition du théorème est satisfaite : les distances issues des données réelles (en l'occurrence, les distances sur les manuscrits) ne sont en général pas arborées. Il est alors usuel de formaliser ce décalage par un problème d'approximation, qui, dans sa version *moindres carrés*, se formule ainsi : étant donné une distance d sur X, déterminer un X-arbre valué minimisant la quantité : $\sum_{(x,y)\in X} [d(x, y) - d_L(f(x), f(y))]^2$. Ce problème est connu pour être NP-difficile (cf. [Day, 1987]), ce qui justifie le recours à des méthodes heuristiques, parmi lesquelles :

- Méthodes de programmation mathématique : Cunningham [1978], Soete [1983] et Brossier [1985].
- Méthodes de réduction : Roux [1988], Gascuel et Lévy [1996].
- Méthodes dites des scores : *ADDTREE* de Sattah et Tversky [1977] et les groupements de Luong [1988].
- Méthodes de regroupement avec parcimonie : NJ (Neighbours Joining) de Saitou ou Nei [1987].

La méthode des Groupements et la méthode NJ ont été utilisées pour les manuscrits sanskrits ; la première pour son adéquation avec le problème posé et la seconde pour sa rapidité et sa stabilité. Elles calculent un X-arbre valué dont on « espère » qu'il représente de façon satisfaisante la distance initiale.

6.3. MÉTHODES DES GROUPEMENTS DE LUONG [1988]

On considère que chaque manuscrit correspond à un sommet de l'arbre à construire, soit une feuille de l'arbre (c'est-à-dire un sommet relié à l'arbre par une seule arête) soit à un nœud (c'est-à-dire un sommet interne de l'arbre qui n'est donc pas une feuille). La méthode part d'une matrice de distance D entre les différents manuscrits. C'est une méthode itérative qui construit des groupements de manuscrits autour d'un nœud qui correspond soit à l'un des manuscrits existants, soit à un manuscrit supposé ou perdu de la tradition textuelle. Pour chaque groupement, on ajoute

des arêtes entre les sommets du groupement et le nœud puis, on supprime de D les sommets du groupement en ne conservant ou en ne rajoutant que le nœud. Les distances entre les sommets restants sont alors calculées et on effectue une nouvelle itération tant qu'il reste au moins trois sommets. On réalise alors un traitement spécifique pour les derniers sommets.

La méthode permet d'obtenir des regroupements de plus de deux sommets, c'est-à-dire des arbres non binaires. Si dans la modélisation phylogénétique, l'ensemble des branchements doivent être binaires, dans le cas des manuscrits, rien ne privilégie la construction d'un stemma bifide. Bédier y voit même là, un des vices de la méthode de Lachmann et développe en 1928 une méthode éditoriale plus légère dans son édition du *Lai de l'Ombre* (cf. [Bédier, 1928]). Le deuxième intérêt de la méthode de Luong est de conduire à des arbres dont tous les sommets étiquetés ne sont pas forcément des feuilles, mais peuvent être des nœuds (on parle alors de *groupement pointé*). La méthode se rapproche donc plus de la modélisation recherchée sur la Figure 11.

On obtient une complexité en $O(n^4)$ pour la méthode classique et en $O(n^5)$ pour les *groupements pointés*.

EXEMPLE. Soient 8 manuscrits ms1, ms2, ..., ms8 dont les dissimilarités d'origine calculées entre les différents manuscrits sont données dans le Tableau 5.

ms1	0							
ms2	32	0						
ms3	48	26	0					
ms4	51	34	42	0				
ms5	50	29	44	44	0			
ms6	48	33	44	38	24	0		
ms7	98	84	92	86	89	90	0	
ms8	148	136	152	142	142	142	148	0
	ms1	**ms2**	**ms3**	**ms4**	**ms5**	**ms6**	**ms7**	**ms8**

Table 5. Matrice des dissimilarités d'origine des manuscrits 1 à 8

La méthode des groupements conduit à une distances d'arbres. On peut alors visualiser les écarts en valeur absolue entre les distances d'arbres obtenues et les dissimilarités d'origine dans le Tableau 5

L'écart absolu moyen est de 2.24 (c'est-à-dire 3 %) sur l'ensemble de l'arbre inféré. À partir de la matrice des distances d'arbres inférées par la méthode des groupements, nous pouvons alors reconstruire l'arbre la la Figure 12.

6.4. MÉTHODE NJ DE SAITOU ET NEI [1987]

C'est une méthode qui combine une approche de distance avec le principe de parcimonie. Les dissimilarités initiales permettent de construire une matrice qui donne un arbre en étoile. Nous partons d'un arbre en étoile dans lequel nous recherchons

	ms1	ms2	ms3	ms4	ms5	ms6	ms7	ms8
ms1		0.6	2.7	1	0.6	2.1	3.3	7.1
ms2	32.6		1.9	0.6	4.2	0.3	0.1	1.7
ms3	45.3	27.9		5.3	1.9	1.4	4.6	1.6
ms4	52	34.6	47.3		0.8	4.7	2.3	0.1
ms5	50.6	33.2	45.9	43.2		0.1	3.5	4.3
ms6	50.1	32.7	45.4	42.7	23.9		2	3.8
ms7	101.3	83.9	96.6	88.3	92.5	92		0.2
ms8	155.1	137.7	150.4	142.1	146.3	145.8	147.8	

Table 6. Matrice des distances d'arbres en dessous de la diagonale et des écarts absolus au-dessus

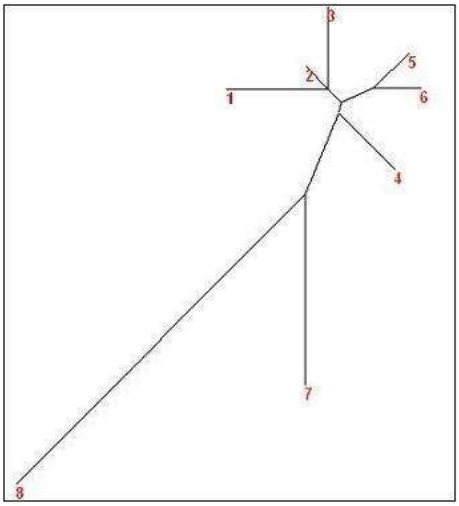


FIGURE 12. Arbre obtenu avec l'algorithme des groupements

le couple de sommets i et j qui, une fois rassemblés, minimise la longueur totale de l'arbre. Lorsque deux sommets sont rassemblés, le noeud représentant leur ancêtre commun est ajouté à l'arbre tandis que les deux feuilles sont enlevées. Ce processus convertit l'ancêtre commun en un noeud terminal dans un arbre de taille réduite dans lequel nous recalculons la matrice de distances. Nous réitérons alors les étapes jusqu'à ce qu'il ne reste plus de sommet dans la matrice.

NJ regroupe les sommets en fonction de leur distance avec l'ensemble des autres sommets, et non pas de leur distance entre eux. Par là-même, NJ minimise aussi la longueur totale des branches.

Si NJ semble moins bien adapté que la méthode des groupements à la modélisation recherchée :

- NJ ne permet d'obtenir que des arbres binaires,
- Les sommets étiquetés sont tous des feuilles de l'arbre,

sa rapidité et son bon comportement font qu'elle demeure une méthode de référence :

- NJ a une complexité en $O(n^3)$.
- Elle permet de retrouver l'arbre si la matrice de distances est correspond à un arbre.
- Elle est considérée comme robuste car, d'une petite variation sur les distances de départ, ne résulte pas dans une mauvaise modélisation d'arbre.

EXEMPLE. En reprenant la matrice utilisée pour les groupements de Luong, on calcule les distances d'arbres avec NJ ainsi que les écarts absolus avec les dissimilarités d'origine dans le Tableau 7 et l'on reconstruit l'arbre de la Figure 13.

ms1		1.9	1.9	1.3	1.3	0.1	0.2	3.6
ms2	33.9		0	0.8	3.2	1.4	2.7	0.9
ms3	46.1	26		3.4	0.4	0.2	7.6	8.3
ms4	49.7	33.2	45.4		2.8	2.6	1.1	1.1
ms5	48.7	32.2	44.4	41.2		0	0.3	1.1
ms6	48.1	31.6	43.8	40.6	24		1.3	0.5
ms7	97.8	81.3	99.6	87.1	89.3	88.7		0
ms8	151.6	135.1	143.7	140.9	143.1	142.5	148	

Table 7. Matrice des distances d'arbres en dessous de la diagonale et des écart absolu au dessus

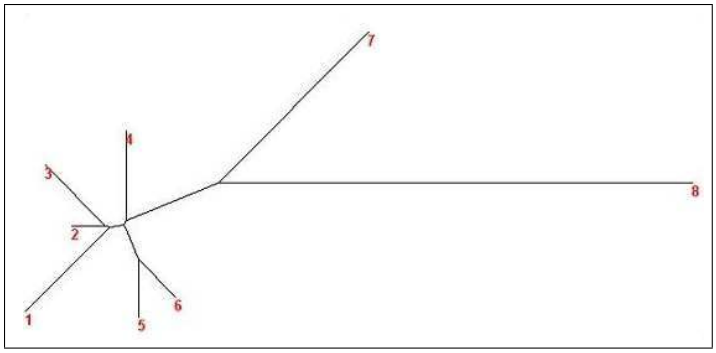


FIGURE 13. Exemple d'arbre obtenu avec l'algorithme NJ

Nous observons que les sommets 1, 2 et 3, qui étaient groupés sur la Figure 12, sont assemblés différemment sur la Figure 13 : le 2 et le 3 sont groupés en premier lieu puis le 1 est groupé avec l'ancêtre de 2 et 3 . Malgré cela, la faible distance entre les ancêtres, suggère finalement la topologie de l'arbre obtenu par les groupements.

L'écart absolu moyen est ici de 1.79 (environ 2.3 %) ce qui correspond à une meilleure approximation que celle obtenue par les groupements. L'approximation par NJ est quasiment toujours meilleure, c'est sans doute le prix à payer pour obtenir une modélisation plus adaptée.

6.5. MANUSCRIT INITIAL

Un arbre *non enraciné* (Figure 14) est une représentation intemporelle des relations entre manuscrits tandis qu'un arbre *enraciné* (Figure 15) spécifie où se situe l'ancêtre commun de tous les manuscrits présents dans l'arbre, considéré comme la racine de l'arbre ou le manuscrit original. La recherche du manuscrit original est un problème important de la méthode stemmatique, et pour tenter de résoudre ce problème, bien que les méthodes des groupements et NJ proposent une racine par construction, nous avons évalué différentes techniques issues de la phylogénétique.

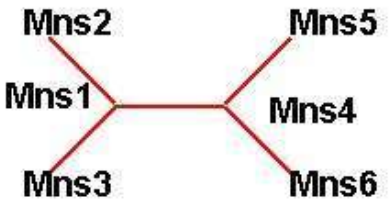


FIGURE 14. Arbre non enraciné

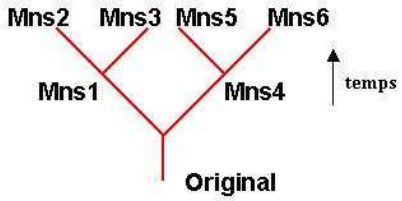


FIGURE 15. Arbre enraciné

UTILISATION DES POINTS PARTICULIERS DES ARBRES

Le centre, la médiane et le n/2 séparateur sont des points particuliers de l'arbre qui servent à déterminer la racine dans des arbres réguliers (cf. Harary [1969]). la Figure 16 est celle d'un arbre régulier. Dans cet exemple, les trois points sont confondus avec la racine. En revanche, sur la Figure 17 qui correspond à une filiation interrompue au niveau du manuscrit 1, les trois points sont toujours confondus mais ne correspondent pas à l'archétype.

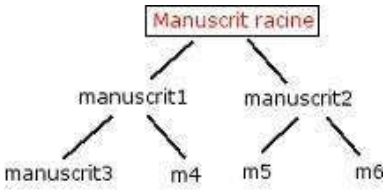


FIGURE 16. Arbre régulier

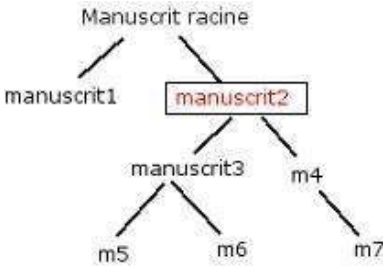


FIGURE 17. Arbre déséquilibré

FABRICATION ARTIFICIELLE D'UN *outgroup*

Pour résoudre le problème en reconstruction phylogénétique, une technique classiquement utilisée est d'ajouter au groupe d'espèces étudiées, une espèce de nature différente appelée *outgroup* (cf. Barriel et Tassy [1998]). Nous en déduisons que la racine est située sur l'arête reliant l'*outgroup* au reste de la phylogénie (cf. Figure 18).

Trois constructions d'*outgroup* différents ont été envisagées sur un corpus fictif dont nous connaissions la racine :

- La construction d'un faux manuscrit constitué de tous les mots du corpus ne donne pas les résultats voulus. En effet, celui-ci a tendance à se rapprocher du manuscrit le plus long (celui qui a le plus de mots communs).

- De même, la construction d'un faux manuscrit constitué de tous les mots communs à tous les manuscrits du corpus ne donne pas de meilleurs résultats. Il se rapproche du manuscrit le plus court.

- Finalement, nous avons utilisé un autre manuscrit complètement différent, mais les résultats ne s'améliorent pas. Le manuscrit se rapproche de celui dont il est le moins loin qui est le manuscrit le plus court.

Actuellement, les diverses expérimentations réalisées pour déterminer un *outgroup* qui convient n'ont pas été concluantes.

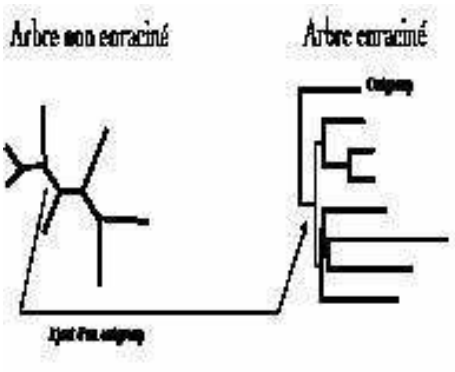


FIGURE 18. Exemple d'outgroup pour enraciner un arbre

7. RÉSULTATS ACTUELS

7.1. RÉSUMÉ DES MÉTHODES DÉVELOPPÉES OU EN COURS DE DÉVELOPPEMENT

Précédemment, nous avons donc envisagé trois méthodes de reconstruction d'arbre à partir de corpus :

1. Alignement de corpus multilingues + Distance d'édition + Méthodes de reconstruction d'arbre basées sur les distances.

2. Alignement de corpus à l'aide du *padapāṭha* + Distance lexicale + Méthodes de reconstruction d'arbre basées sur les distances.

3. Distance de compression + Méthodes de reconstruction d'arbre basées sur les distances.

7.2. CORPUS DE TEST POUR LES MÉTHODES

Les algorithmes sont testés sur différents corpus afin de vérifier et expérimenter les méthodes mises au point.

Pour les premiers tests, plusieurs corpus « fictifs » ont été réalisés : partant d'un premier texte considéré comme l'original, nous dégradons celui-ci « à la façon d'un copiste » pour obtenir tout un corpus du même texte plus ou moins dégradé. Nous avons alors l'avantage de connaître le stemma lié à notre corpus et nous pouvons aussi réaliser aisément certain corpus ayant des stemmata particuliers (exemple : arbre déséquilibré par perte de manuscrits dans une branche). L'aspect négatif est bien entendu que nous ne maîtrisons pas les techniques de copie d'un manuscrit réalisé sur des siècles, dans des graphies différentes par de nombreux copistes. Lors de l'expérimentation, les méthodes se comportent en général très bien et elle permet de les affiner.

Par la suite, nous avons aussi testé les programmes sur un corpus de *Chain Letter* collecté par Bennett [to appear]. Une lettre-chaîne est une lettre demandant au destinataire d'en envoyer à son tour de multiples copies, de telle manière que sa propagation s'accroisse de façon exponentielle. Elles sont rédigées dans l'espoir d'apporter bonheur, argent... Le destinataire se transforme alors en un copiste comme pour une transmission de manuscrits. Les résultats permettent de retrouver la classification des lettres telle qu'elle a été effectuée par Bennett. En effet sur les Figures 19 et 20, on observe que les sous ensembles principaux représentés par les formes géométriques sont conservés.

Enfin, un travail est actuellement en cours sur une édition critique pour laquelle le stemma codicum a déjà été réalisé par des philologues. Le but étant de comparer les résultats obtenus par les philologues et ceux des méthodes précédemment construites.

7.3. APPLICATION SUR DES MANUSCRITS

La reconnaissance des mots à travers le *padapāṭha* n'étant actuellement pas terminée, seules les méthodes avec la distance de Levensthein et la distance de compression ont été testées sur le corpus des manuscrits.

Pour les comparer, on a réalisé 3 corpus d'une cinquantaine de manuscrits :

- Un corpus n° 1 contenant les 34 premiers paragraphes du premier chapitre.
- Un corpus n° 2 contenant les 17 premiers paragraphes du premier chapitre.
- Un corpus n° 3 contenant les 17 derniers paragraphes du premier chapitre.

L'expérimentation fait rapidement apparaître les manuscrits qui sont très délabrés car ceux-ci se situent très loin des autres dans le sens vertical comme le manuscrit io5 des Figures 21 ou 22.

Les racines de l'arbre sur les Figures 21 et 22 sont celles proposées par la méthode NJ dans le cadre de son algorithme. Les méthodes proposées au paragraphe 6.5. n'ont actuellement pas donné satisfaction et l'on n'a pas réussi à déterminer la racine (le manuscrit original) par une méthode extérieure intéressante.

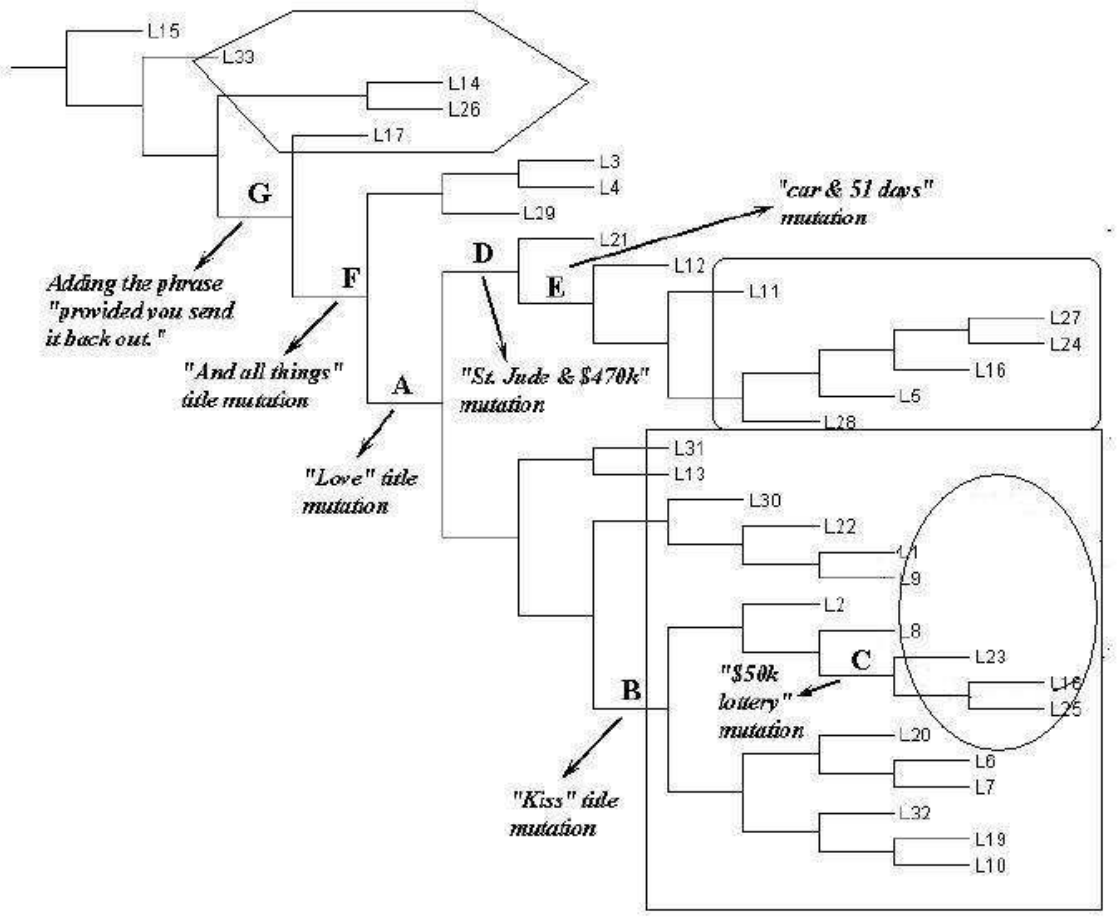


FIGURE 19. Arbre établi par Bennett à partir du corpus de “Chain Letter”

Dans notre expérimentation, nous ne nous sommes pas préoccupés de vérifier la stabilité de l’arbre obtenu. Par exemple, l’introduction d’un nouveau manuscrit ou son enlèvement donnera-t-il un arbre similaire, c’est-à-dire dont la topologie sera identique ? Actuellement, des résultats similaires sont obtenus par les différentes méthodes et sur les trois corpus ce qui tend à confirmer la stabilité de la méthode. De plus, l’analyse des premiers résultats confiée à des sanskritistes permet de confirmer l’intérêt des techniques employées. En effet, sur les Figures 21 et 22, les lettres A B C et D sont celles de groupements de manuscrits réalisés par les sanskritistes. Nous constatons qu’ils s’accordent avec les classifications obtenues.

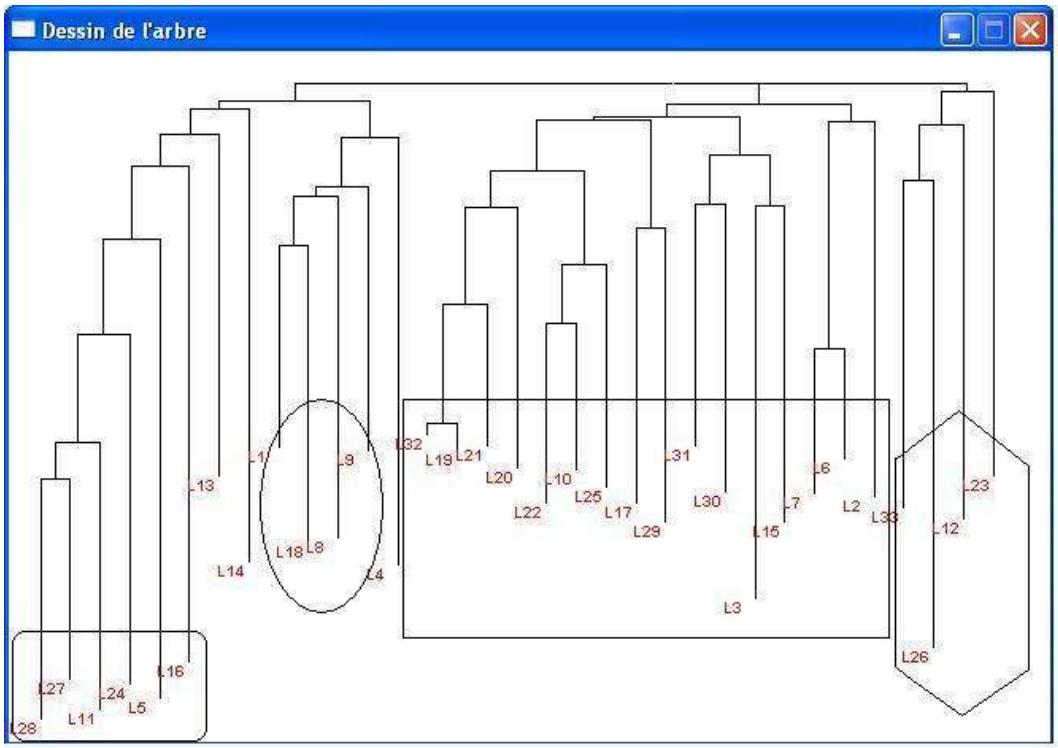


FIGURE 20. Arbre obtenu par NJ à partir du corpus de “Chain Letter”

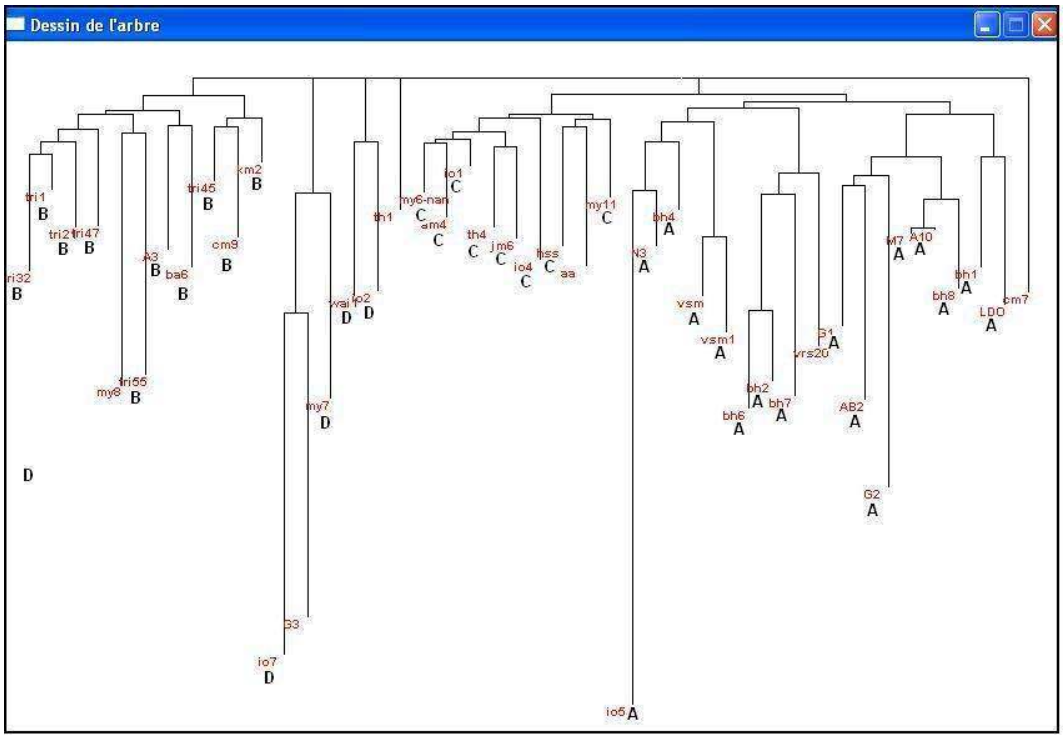


FIGURE 21. Arbre Hiérarchique avec distance de Levensthein

Une autre expérimentation qui a été réalisée, consiste à lister sur notre arbre les différentes graphies des manuscrits. Seul le *malayālam* se trouve vraiment isolé dans

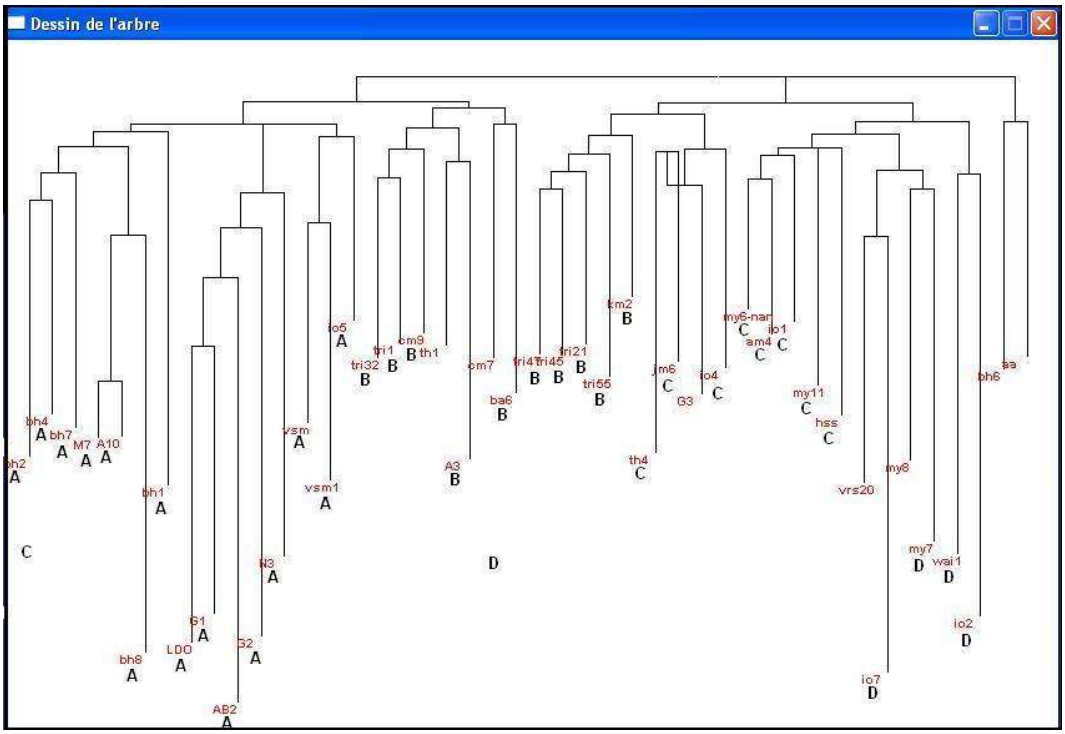


FIGURE 22. Arbre Hiérarchique avec distance de compression

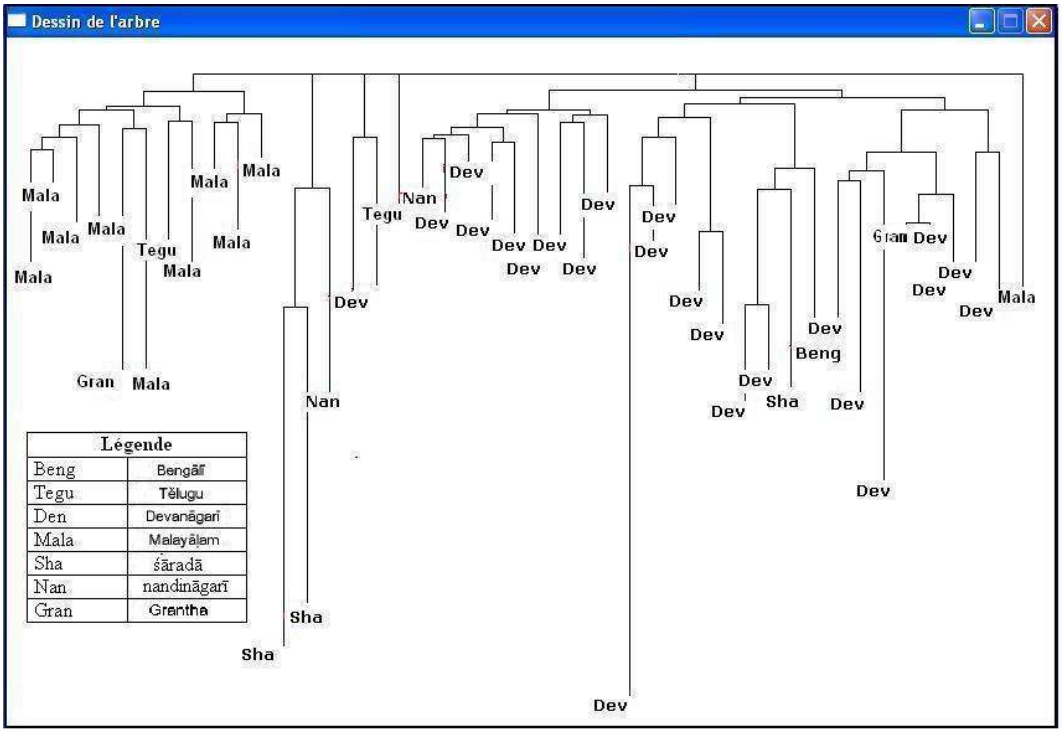


FIGURE 23. Arbre Hiérarchique avec les graphies des manuscrits

notre arbre. L'interprétation, pour les autres graphies, étant plutôt le travail des sanskritistes, on peut néanmoins se douter que les filiations se créent de préférence intra-graphie que inter-graphie.

8. CONCLUSIONS ET PERSPECTIVES

À PROPOS DE LA STRUCTURE DE REPRÉSENTATION

Nous pouvons avant tout nous demander si, après le nombre important de prétraitements effectués sur le corpus et si, par le choix d'une distance, nous classons toujours la filiation des manuscrits et non pas les graphies, les écoles de copistes ou bien les dégradations du temps constatées sur les manuscrits, etc.

Pour valider notre classification, le nombre de méthodes et de corpus permettent d'obtenir à chaque fois un nouvel arbre. Il apparaît difficile d'exploiter la vingtaine d'arbres obtenus. Il est alors souhaitable de comparer et d'inférer un nouvel arbre par des techniques de consensus ou autres.

Un autre problème est de savoir déterminer au mieux la racine. Actuellement les techniques utilisées n'ont pas donné satisfaction. Pour cela il faut sûrement intégrer des « informations extérieures » comme la datation. Il est alors nécessaire de construire notre arbre en y incorporant au fur et à mesure ces connaissances supplémentaires.

Dans ce cadre, il est sûrement intéressant d'envisager la reconstruction de l'arbre par des méthodes d'intermédiarité. Il faut pour cela affiner le concept de manuscrit intermédiaire et peut-être, à l'aide de l'alignement multiple sur les mots, définir un score d'intermédiarité.

On peut aussi penser que la structure d'arbre doit être remise en cause au profit d'une structure de graphe. En effet la structure composite[11] de certains manuscrits ou la contamination[12] contredisent le modèle phylogénétique puisqu'un manuscrit peut provenir de la filiation (copie) de deux manuscrits différents. Il apparaît alors intéressant de construire un indice de contamination de notre corpus au-delà duquel l'arbre n'est pas valide et peut être même le stemma (cf. [Mass, 1927]).

AU SUJET DES INFORMATIONS EXTÉRIEURES AUX TEXTES

Une étude des règles de l'acte de copie peut permettre d'orienter l'arbre ou le graphe. Nous pouvons étudier statistiquement les usages des copistes (évolution statistique des mots dans le temps, modification des caractères selon les écoles de scribes ...).

Les études codicologiques et paléographiques des manuscrits peuvent apporter des confirmations au niveau de notre classification, voire la modifier.

Enfin des connaissances supplémentaires sur le sanskrit peuvent être utilisées pour passer d'une analyse au niveau des mots à une analyse au niveau des lemmes sans doute plus riche de sens.

Remerciement. Je tiens à remercier M. Del Vigna pour ses conseils, sa rigueur et sa patience lors de la mise en forme de cet article.

[11]Composite : se dit d'un manuscrit composé de parties de dates et d'origines diverses (syn. : hétérogène).

[12]Contamination : copie d'un manuscrit sur plusieurs modèles (syn. : corruption, hybridation).

BIBLIOGRAPHIE

BARRIEL V., TASSY P. (1998), "Rooting with multiple outgroup: Consensus versus parsimony", *Cladistics* 14(2), p. 193-200.

BARTHÉLEMY J.-P., GUÉNOCHE A. (1988), *Les arbres et les représentations des proximités*, Paris, Masson.

BÉDIER J. (1928), « La tradition manuscrite du Lai de l'Ombre. Réflexions sur l'art d'éditer les anciens textes », *Romania* 54, p. 162-186, p. 321-356.

BELLMAN R.E. (1957), *Dynamic programming*, Princeton (New Jersey), Princeton University Press.

BENNETT C., LI M., MA B., "Linking chain letters", *Scientific American*, [to appear].

BENNETT C.H., GACS P., LI M., VITANYI P., ZUREK W.H. (1998), "Information distance", *IEEE Transactions on information theory* 44, p. 1407-1423.

BEVENOT M.S.J. (1961), *The tradition of manuscripts: a study in the transmission of St Cyprian's treatrises*, Oxford, Clarendon Press.

BROSSIER G. (1985), « Approximation des dissimilarités par des arbres additifs », *Mathématiques et Sciences humaines* 91, p. 5-31.

BUNEMAN P. (1971(a)), "Filiations of manuscripts", F.R. Hodson and D.G. Kendall (eds), *Mathematics in Archaeological and Historical Sciences*, Edinburgh, Edinburgh University Press.

BUNEMAN P. (1971(b), "The recovery of trees from measures of dissimilaritie", D.G. Kendall and P. Tautu (eds.), *Mathematics in Archeological and Historical Sciences*, Edinburgh, Edinburgh University Press, p. 387-395.

BURROWS M., WHEELER D.J. (1994), "A block-sorting lossless data compression algorithm", *SRC Research report 124, Digital systems research center*, Palo Alto.

CILIBRASI R., VITÁNYI P. (2005), "Clustering by compression", *IEEE Transactions on information theory* 51(4), p. 1523-1545.
URL http://www.cwi.nl/paulv/papers/cluster.pdf

CSERNEL M., BERTRAND P. (2005), « Comparaison de manuscrits sanskrits », *Modulad* 33(1), p. 1-20, [édition critique de classification].

CUNNINGHAM J.P. (1978), "Free trees and bidirectional trees as representations of psychological distance", *Journal of mathematical psychology* 17, p. 165-188.

DAY W.H.E. (1987), "Computational complexity of inferring phylogenies from dissimilarity matrices", *Bulletin of Mathematical Biology* 49, p. 461-467.

FILLIOZAT J. (1941), *Catalogue du fonds sanskrits*, Paris, Adrien Maisonneuve édition.

GALE W.A., CHURCH K.W. (1991), "A program for aligning sentences in bilingual corpora", *Comutational linguistics* 19, p. 75-102.

GASCUEL O., LÉVY D. (1996), "A reduction algorithm for approximating a (nonmetric) dissimiarity by a tree distance", *Journal of Classification* 13, p. 129-155.

GRIFFITH I.G. (1969), "Numerical taxonomy and some primary manuscripts of the gospels", *JTS* 20, p. 389-406.

GRÜWALD P., VITÁNYI P. (2004), "Shannon information and Kolmogorov complexity", *CoRR*, cs.IT/0410002.

HAMMING R.W. (1950), "Error detecting and error correcting codes", *Bell System Technical Journal* 26(2), p. 147-160.

HARARY F. (1969), *Graph theory*, Reading (Mass.), Addison-Reading.

HUFFMAN D.A. (1952), "A method for the construction of minimum redundancy codes", *Proceedings of the Institut of Radio Engineers* 40(9), p. 1098-1101.

KOLMOGOROV A.N. (1965), "Three approaches to the quantitative definition of information", *Prob. Inf. Trans.* 1, p. 1-7.

LEMPEL A., ZIV J. (1977), "A universal algorithm for sequential data compression", *IEEE Transactions on information theory* 23(3), p. 337-343.

LEVENSHTEIN V.I. (1966), "Binary codes capable of correcting deletions, insertions and reversals", *Soviet Physics-Doklandy* 10(8), p. 707-710.

LUONG X. (1988), *Méthodes d'analyse arborée. Algorithmes. Applications*, thèse de doctorat, Université Paris V.

MAAS P. (1927), *Textkritik*, Leipzig und Berlin, B.G. Teubner.

QUENTIN H. (1926), *Essais de critique textuelle*, Paris, Picard.

RASHED M. (2005), *De la génération et la corruption d'Aristote*, Les Belles Lettres.

ROUX M. (1988), "Techniques of approximation for building two tree structures", *Recent developments in clustering and data analysis*, New York, Academic Press.

SAITOU N., NEI M. (1987),"The neighbor-joining method: a new method for reconstructing phylogenetic trees", *Mol. Biol. Evol.* 4, p. 406-425.

SALEMANS B.J.P. (2000), *Building stemmas with the computer in a cladistic, neo-lachmannian way: the genealogy of the fourteen versions of Lanseloet van Denemerken*, thèse de doctorat, Nijmegen (The Netherlands), Katholieke Universiteit Nijmegen, 351 pages.

SATTAH S., TVERSKY A. (1977), "Additive similarity", *Psychometrika* 42, p. 319-345.

SHANNON C.E. (1948), "A mathematical theory of commuication", *Bell. Sys. Tech. J.* 27, p. 379-423, p. 623-656.

SOETE (de) G. (1983), "Least squares algorithma for additive trees to proximity data", *Psychometrika* 48, p. 621-626.

VERTHUIS F. (1991), *Package devanagari pour Tex Latex.*
URL ftp://ftp.tex.ac.uk/tex-archive/language/devanagari/velthuis/

WAGNER A.R., FISHER J.M. (1974), "The string to string correction problem", *Journal of the Association for Computing Machinery* 21(1), p. 168-173.

WATERMAN M.S. (1995), *Introduction to computational biology*, London, Chapman and Hall.